Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau.

Alain Durmus, Éric Moulines, Marcelo Pereyra

Sampling from convex non continuously differentiable functions, when Moreau meets Langevin

Alain Durmus[1] Éric Moulines [2] Marcelo Pereyra [3]

February 3, 2016

Keywords: Markov Chain Monte Carlo, Metropolis Adjusted Langevin Algorithm, Rate of convergence

AMS subject classification (2010): primary 65C05, 60F05, 62L10; secondary 65C40, 60J05,93E35

Abstract

In this paper, two new algorithms to sample from possibly non-smooth log-concave probability measures are introduced. These algorithms use Moreau-Yosida envelope combined with the Euler-Maruyama discretization of Langevin diffusions. They are applied to a deconvolution problem in image processing, which shows that they can be practically used in a high dimensional setting. Finally, non-asymptotic bounds for one of the proposed methods are derived. These bounds follow from non-asymptotic results for ULA applied to probability measures with a convex continuously differentiable log-density with respect to the Lebesgue measure.

1 Introduction

Sampling for high-dimensional distribution, known up to a normalizing constant, is the crux for Bayesian inference. Assume that we are willing to sample a distribution with density π with respect to the Lebesgue measure on $\mathbb{R}^d$ of the form $x \mapsto \mathrm{e}^{-U(x)} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y$, for some measurable function $U : \mathbb{R}^d \to (-\infty, +\infty]$, which satisfies the following condition.

H 1. *$U = f + g$, where $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to (-\infty, +\infty]$ are two lower bounded, convex functions satisfying:*

(i) f is continuously differentiable and gradient Lipschitz with Lipschitz constant L_f, i.e. for all $x, y \in \mathbb{R}^d$

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\| \ . \tag{1}$$

(ii) g is lower semi-continuous and $\int_{\mathbb{R}^d} \mathrm{e}^{-g(y)} \mathrm{d}y \in (0, +\infty)$.

[1]LTCI, Telecom ParisTech 46 rue Barrault, 75634 Paris Cedex 13, France. alain.durmus@telecom-paristech.fr
[2]Centre de Mathématiques Appliquées, UMR 7641, Ecole Polytechnique, France. eric.moulines@polytechnique.edu
[3]Department of Mathematics, University of Bristol, University Walk, Clifton, Bristol BS8 1TW, U.K. marcelo.pereyra@bristol.ac.uk

In the Bayesian setting, f is the opposite of the log-likelihood. A typical example is the regression model where $f(x) = \|y - \mathrm{M}x\|^2$, where $y \in \mathbb{R}^N$ are the responses (N being the number of observations), $\mathrm{M} \in \mathbb{R}^{d \times N}$ is a known regression matrix, and x are the regression parameters. The function g is the potential of a prior distribution. Since we willing to cover classes of prior which favors sparsity, we do not assume that g is differentiable. For example, taking a Laplace prior $g(x) = \|x\|_1$ in a regression problem leads to the Bayesian LASSO regression; see Park and Casella [2008].

If U is continuously differentiable on $\mathbb{R}^d$, the Langevin stochastic differential equations (SDE) associated with π is given by

$$\mathrm{d}\mathbf{X}_t^L = -\nabla U(\mathbf{X}_t^L)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t^d \;, \tag{2}$$

where $(B_t^d)_{t \geq 0}$ is a d-dimensional Brownian motion. Under mild assumptions, this equation has a unique strong solution; the semi-group associated with the Langevin SDE is reversible and ergodic with respect to π which is hence stationary, see Khas'minskii [1960]. To sample π, this suggests to consider the Markov chain given by the Euler-Maruyama discretization of (2), defined given the current state X_k by

$$X_{k+1} = X_k - \gamma_{k+1}\nabla U(X_k) + \sqrt{2}Z_{k+1} \;, \tag{3}$$

where $(\gamma_k)_{k \geq 1}$ is a nonincreasing sequence of stepsizes and $(Z_k)_{k \geq 1}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables. This scheme has been first introduced in molecular dynamics by Ermak [1975] and Parisi [1981], and then popularized in the machine learning community by Grenander [1983], Grenander and Miller [1994] and computational statistics by Neal [1993] and Roberts and Tweedie [1996]. Following Roberts and Tweedie [1996], this algorithm is referred to as the Unadjusted Langevin Algorithm (ULA). Under additional assumptions on U and if $\lim_{k \to +\infty} \gamma_k = 0, \sum_{k=0}^{\infty} \gamma_k = +\infty$, it has been shown in Lamberton and Pagès [2002], Lemaire [2005] that for all h in an appropriately defined class of functions

$$\lim_{n \to +\infty} \left(\sum_{i=1}^{n} \gamma_i \right)^{-1} \sum_{k=0}^{n} \gamma_{k+1} h(X_k) = \int_{\mathbb{R}^d} h(y)\mathrm{d}\pi(y) \;.$$

This result was later extended in Durmus and Moulines [2015], which provides non-asymptotic deviation bounds. Under weak additional conditions, a central limit theorem can be obtained.

For constant stepsizes $\gamma_k = \gamma$ for all k, the Markov chain associated to the Euler-Maruyama discretization also converges (under weak additional technical conditions) to a probability measure π_γ, which is different from π. Dalalyan [2014] and Durmus and Moulines [2015] give non-asymptotic bounds for the total variation between these two probability measures with explicit dependence on the stepsize γ and the dimension d. To get a reversible Markov chain with respect to π, and therefore drop the asymptotic bias, Rossky et al. [1978] and Roberts and Tweedie [1996] have suggested to include the one-step transition kernel of the Euler-Maruyama discretization (3) as a proposal kernel in a Metropolis-Hastings algorithm. Because π is this time the target distribution, this algorithm is called the Metropolis Adjusted Langevin Algorithm (MALA), a term coined by Roberts and Tweedie [1996].

As mentioned earlier, the main motivation of this paper is to provide efficient methods to sample high-dimensional regression model with sparsity inducing prior g, which satisfies **H**1-(ii). In the Bayesian linear regression with Laplace prior for example, U is differentiable almost everywhere, and being a proper convex function, has subgradient everywhere. Therefore, whereas U is not differentiable, the ULA and MALA algorithms could still be formally applied. This solution is not satisfactory. First, some technical difficulties arise however when defining the associated

SDE, which no longer has strong solutions. Second, experimental evidence shows however that the MALA algorithm mixes poorly, see [Pereyra, 2015, Figure 4]. In some applications, ∇U is not even well defined on a subset of $\mathbb{R}^d$ with non null Lebesgue measure. This problem occurs when sampling a distribution supported on on a bounded convex set $\mathcal{K}$. In such case, the potential g is bounded on $\mathcal{K}$ and infinite outside $\mathcal{K}$.

On the other hand, new efficient first-order optimization algorithms have been recently introduced in the convex optimization literature to compute a Maximum a posteriori of such non-smooth convex potential; see Parikh and Boyd [2013] and the references therein. These optimisation algorithms use gradient descent for the smooth part of the the potential combined with a proximal step for the non-smooth part.

In this paper, we introduce two new algorithms designed to sample from $\pi = \mathrm{e}^{-U}$ where the potential U satisfies **H**1. The idea is to construct an convex and smooth distribution using the Moreau-Yosida proximal operator, and then apply either the ULA or MALA algorithm to these approximations. Second, to compute expectation under the target distribution π, an importance sampling step is introduced.

The paper is organized as follows. In Section 3, the Moreau-Yosida regularization is introduced and the properties of the regularized target distribution is presented. In Section 3 non-asymptotic bounds in total variation for the ULA algorithms are presented, since it is the first step of one of our algorithms. To illustrate our findings, a limited Monte Carlo experiment is presented in Section 4. We consider a high-dimensional Bayesian linear inverse problem to illustrate the practical feasibility and validity of the proposed algorithms in a realistic context.

Notations and Conventions

Denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. For all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$, denote by $\mathrm{Vol}(\mathsf{A})$ its Lebesgue measure. Denote by $\mathbb{M}(\mathbb{R}^d)$ the set of all Borel measurable functions on $\mathbb{R}^d$ and for $f \in \mathbb{M}(\mathbb{R}^d)$, $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $f \in \mathbb{M}(\mathbb{R}^d)$ a μ-integrable function, denote by $\mu(f)$ the integral of f w.r.t. μ. For two probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, the total variation of μ and ν is defined as

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{f\in\mathbb{M}(\mathbb{R}^d), \|f\|_\infty \le 1} \left| \int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x) - \int_{\mathbb{R}^d} f(x)\mathrm{d}\nu(x) \right| \tag{4}$$

Let $f : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper function, denote by $\partial f(x)$ the subdifferential of f at $x \in \mathbb{R}^d$. If f is a Lipschitz function, namely there exists $C \ge 0$ such that for all $x, y \in \mathbb{R}^d$, $|f(x) - f(y)| \le C \|x - y\|$, then denote $\|f\|_{\mathrm{Lip}} = \inf\{|f(x) - f(y)| \|x - y\|^{-1} \mid x, y \in \mathbb{R}^d, x \ne y\}$. For $k \ge 0$, denote by $C^k(\mathbb{R}^d)$, the set of k-times continuously differentiable functions. For $f \in C^2(\mathbb{R}^d)$, denote by Δf the Laplacian of f. Denote for all $\mathrm{q} \ge 1$, the ℓ_q norm $|\cdot|_\mathrm{q}$ on $\mathbb{R}^d$ by for all $x \in \mathbb{R}^d$, $\|x\|_\mathrm{q} = (\sum_{i=1}^d |x_i|^\mathrm{q})^{1/\mathrm{q}}$. For all $x \in \mathbb{R}^d$ and $M > 0$, denote by $\mathrm{B}(x, M)$, the ball centered at x of radius M. For a closed convex $\mathcal{K} \subset \mathbb{R}^d$, denote by $\mathrm{proj}_\mathcal{K}(\cdot)$, the projection onto $\mathcal{K}$, and $\iota_\mathcal{K}$ the convex indicator of $\mathcal{K}$ defined by $\iota_\mathcal{K}(x) = 0$ if $x \in \mathcal{K}$, and $\iota_\mathcal{K}(x) = +\infty$ otherwise. For all subsets $\mathsf{E}_1, \mathsf{E}_2 \subset \mathbb{R}^d$, denote by $\mathsf{E}_1 + \mathsf{E}_1 = \{x + y \mid x \in \mathsf{E}_1, \ y \in \mathsf{E}_2\}$. In the sequel, we take the convention that $\inf \emptyset = \infty$, $1/\infty = 0$ and for $n, p \in \mathbb{N}$, $n < p$ then $\sum_p^n = 0$ and $\prod_p^n = 1$. Denote by $\mathbf{\Phi}$ and $\mathbf{\Phi}^{-1}$ the cumulative distribution function and the quantile function of the a standard Gaussian variable.

2 Moreau-Yosida Regularized Langevin Algorithms: MYULA and MYMALA

The key tools in this work are proximal operators and Moreau-Yosida envelopes; see Parikh and Boyd [2013] and Polson et al. [2015]. Let $\mathrm{h} : \mathbb{R}^d \to (-\infty, +\infty]$ be a l.s.c convex function and $\lambda > 0$. The λ-Moreau-Yosida envelope $\mathrm{h}^\lambda : \mathbb{R}^d \to \mathbb{R}$ and the proximal operator $\mathrm{prox}_{\mathrm{h}}^\lambda : \mathbb{R}^d \to \mathbb{R}^d$ associated with h (see [Rockafellar and Wets, 1998, Chapter 1 Section G]) are defined for all $x \in \mathbb{R}^d$ by

$$\mathrm{h}^\lambda(x) = \inf_{y \in \mathbb{R}^d} \left\{ \mathrm{h}(y) + (2\lambda)^{-1} \|x - y\|^2 \right\} \leq \mathrm{h}(x) \,, \tag{5}$$

For every $x \in \mathbb{R}^d$, the minimum is achieved at a unique point, $\mathrm{prox}_{\mathrm{h}}^\lambda(x)$, which is characterized by the inclusion

$$x - \mathrm{prox}_{\mathrm{h}}^\lambda(x) \in \gamma \partial \mathrm{h}(\mathrm{prox}_{\mathrm{h}}^\lambda(x)) \,. \tag{6}$$

The Moreau-Yosida envelope is a regularized version of g, which approximates g from below. The proximal (or proximity) operator specifies the (unique) point solving the optimization problem. The parameter λ defines a trade-off between the two objectives of minimizing g and staying close to x. Since for $0 < \lambda < \lambda'$, $\left\{ \mathrm{h}(y) + (2\lambda')^{-1} \|x - y\|^2 \right\} \leq \left\{ \mathrm{h}(y) + (2\lambda)^{-1} \|x - y\|^2 \right\}$ for all $x, y \in \mathbb{R}^d$, we get $\mathrm{h}^{\lambda'}(x) \leq \mathrm{h}^\lambda(x)$. In addition, as $\lambda \downarrow 0$, converges h^λ converges pointwise h, *i.e.* for all $x \in \mathbb{R}^d$,

$$\mathrm{h}^\lambda(x) \uparrow \mathrm{h}(x) \,, \quad \text{as } \lambda \downarrow 0 \,. \tag{7}$$

Furthermore, the function h^λ is convex and continuously differentiable with gradient given by

$$\nabla \mathrm{h}^\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_{\mathrm{h}}^\lambda(x)) \,. \tag{8}$$

The proximal operator behaves like a gradient-descent step for the function h in the sense that $\mathrm{prox}_{\mathrm{h}}^\lambda(x) = x - \lambda \nabla \mathrm{h}^\lambda(x)$. The proximal operator is a monotone operator [Rockafellar and Wets, 1998, Proposition 12.19], *i.e.* for all $x, y \in \mathbb{R}^d$,

$$\left\langle \mathrm{prox}_{\mathrm{h}}^\lambda(x) - \mathrm{prox}_{\mathrm{h}}^\lambda(y), x - y \right\rangle \geq 0 \,, \tag{9}$$

which implies that the Moreau-Yosida envelope is smooth in the sense that its gradient is Lipshitz: $\left\| \nabla \mathrm{h}^\lambda(x) - \nabla \mathrm{h}^\lambda(y) \right\| \leq \lambda^{-1} \|x - y\|$, for all $x, y \in \mathbb{R}^d$.

For example, the proximal operator associated with the ℓ_1-norm and the parameter λ on $\mathbb{R}^d$ is the soft thresholding operator defined by for all $x \in \mathbb{R}^d$, $\mathrm{prox}_{|\cdot|_1}^\lambda(x)$ is the d-dimensional vector whose component $i \in \{1, \cdots, d\}$ is equal to $\mathrm{sign}(x_i) \times \max(|x_i| - \lambda, 0)$ (see e.g. [Parikh and Boyd, 2013, Section 6.5.2]). In the case where $g = \iota_{\mathcal{K}}$ for a closed convex subset $\mathcal{K} \subset \mathbb{R}^d$, the proximal operator is simply the projection onto $\mathcal{K}$ for all $\lambda > 0$.

Under **H**1, if g is not differentiable, but the proximal operator associated with g is available, its λ-Moreau Yosida envelope g^λ can be considered. This leads to the approximation of the potential $U^\lambda : \mathbb{R}^d \to \mathbb{R}$ defined for all $x \in \mathbb{R}^d$ by

$$U^\lambda(x) = f(x) + g^\lambda(x) \,.$$

The following proposition implies that the function e^{-U^λ} can be renormalized to define a density of a probability measure π^λ on $\mathbb{R}^d$.

Proposition 1. *Assume **H**1. Then for all $\lambda > 0$, it holds $0 < \int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(y)} \mathrm{d}y < +\infty$.*

Proof. The proof is postponed to Section 5.1. □

Under **H**1 and using (5), U^λ is continuously differentiable and gradient Lipschitz. Given a regularization parameter $\lambda > 0$ and a sequence of stepsizes $\{\gamma_k,\ k \in \mathbb{N}^*\}$, the algorithm produces the Markov chain $\{X_k^{\mathrm{M}},\ k \in \mathbb{N}\}$: for all $k \geq 0$

$$X_{k+1}^{\mathrm{M}} = X_k^{\mathrm{M}} - \gamma_{k+1} \left\{ \nabla f(X_k^{\mathrm{M}}) + \lambda^{-1}(X_k^{\mathrm{M}} - \mathrm{prox}_g^\lambda(X_k^{\mathrm{M}})) \right\} + \sqrt{2\gamma_{k+1}} Z_{k+1} \,, \tag{10}$$

where $\{Z_k,\ k \in \mathbb{N}^*\}$ is a sequence of i.i.d. d dimensional standard Gaussian random variables. Since, as shown in Durmus and Moulines [2015], the ULA algorithm can be used to target π^λ (provided that the sequence of stepsize $\{\gamma_k,\ k \in \mathbb{N}\}$ decreases to zero at an appropriate rate), the algorithm (10) target the smoothed distribution π^λ. Hence, to compute the expectation of a function $h : \mathbb{R}^d \to \mathbb{R}$ under π from $\{X_k^{\mathrm{M}}\ ;\ 0 \leq k \leq n\}$, an importance sampling scheme is used to correct the smoothing. This step amounts to approximate $\int_{\mathbb{R}^d} h(x)\pi(x)\mathrm{d}x$ by the weighted sum

$$\mathrm{S}_n^h = \sum_{k=0}^n \omega_{k,n} h(X_k) \,, \text{ with } \omega_{k,n} = \left\{ \sum_{k=0}^n \gamma_k \mathrm{e}^{\bar{g}^\lambda(X_k^{\mathrm{M}})} \right\}^{-1} \gamma_k \mathrm{e}^{\bar{g}^\lambda(X_k^{\mathrm{M}})} \,, \tag{11}$$

where for all $x \in \mathbb{R}^d$

$$\bar{g}^\lambda(x) = g^\lambda(x) - g(x) = g(\mathrm{prox}_g^\lambda(x)) - g(x) + (2\lambda)^{-1} \left\| x - \mathrm{prox}_g^\lambda(x) \right\|^2 \,.$$

This algorithm will be called the *Moreau-Yosida Unadjusted Langevin Algorithm* (MYULA).

Note that if the stepsize $\gamma_k = \gamma$ is constant, the ULA sequence $\{X_k^{\mathrm{M}},\ k \in \mathbb{N}\}$ does not longer target π^λ, but a distribution π_γ^λ which depends on the stepsize (note that the approximation error $\|\pi^\lambda - \pi_\gamma^\lambda\|_{\mathrm{TV}} = \mathcal{O}(\gamma^{1/2})$ - see Proposition 5). To remove this bias, we can add an Hastings-Metropolis step, which will produce a Markov chain $\{\tilde{X}_k^\lambda,\ k \in \mathbb{N}\}$ which is reversible this time with respect to π^λ and use similarly an importance sampling step to correct for the bias introduced by smoothing. This algorithm will be called the *Moreau-Yosida Regularized Metropolis Adjusted Langevin Algorithm* (MYMALA).

To justify, the importance sampling step, we give in the following some results and bounds on the behaviour of $\|\pi^\lambda - \pi\|_{\mathrm{TV}}$ in function of λ.

Proposition 2. *Assume **H**1.*

(a) Then, $\lim_{\lambda \to 0} \|\pi^\lambda - \pi\|_{\mathrm{TV}} = 0$.

(b) Assume in addition that g is Lipschitz. Then for all $\lambda > 0$,

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq \lambda \|g\|_{\mathrm{Lip}}^2 \,.$$

(c) If $g = \iota_{\mathcal{K}}$ where $\mathcal{K}$ is a convex body of $\mathbb{R}^d$. Then for all $\lambda > 0$ we have

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq 2 \left(1 + \mathsf{D}(\mathcal{K}, \lambda)\right)^{-1} \,, \tag{12}$$

where $\mathsf{D}(\mathcal{K}, \lambda)$ is explicit in the proof, and is of order $\mathcal{O}(\lambda^{-1})$ as λ goes to 0.

Proof. The proof is postponed to Section 5.2. □

3 Main results

In this section, we fix the regularization parameter $\lambda > 0$ and for ease of notation denote by V and μ, U^λ and π^λ respectively. We apply in MYULA, a ULA step to sample from the distribution μ having a potential V which satisfies the following conditions.

H2 (V). *i) The function $V : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, convex and gradient Lipschitz with constant L_c.*

ii) There exist $\rho_c > 0$ and $R_c \geq 0$ such that for all $x \in \mathbb{R}^d$, $\|x - x^\star\| \geq R_c$,

$$V(x) - V(x^\star) \geq \rho_c \|x - x^\star\| \ , \tag{13}$$

where $x^\star$ is a minimizer of V.

It has been observed that g^λ is λ^{-1}-gradient Lipschitz, which implies that a upper bound for L_c is $L_f + \lambda^{-1}$. However, this bound strongly depends on the decomposition of U, which is arbitrary, so can be pessimistic. For instance, if U is continuously differentiable, g can be chosen to be 0 which implies $V = U$ and $L_c = L_f$. Furthermore, **H**2-ii) holds for V since by Lemma 6 and Proposition 1 there exists $C_1, C_2 > 0$ such that $V(x) \geq C_1 \|x\| - C_2$, which easily implies (13) with $\rho_c \leftarrow C_1/2$ and $R_c \leftarrow 2(C_2 + \|x^\star\| + V(x^\star))/C_1$. But these constants are non-quantitative and that is why we assume they can be estimated. To get non-asymptotic bounds for MYULA, we are interested in this part to get non-asymptotic to ULA applied to the probability measure μ.

For $\gamma > 0$, consider the Markov kernel R_γ associated to the Euler-Maruyama discretization (10) is given, for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$ by

$$R_\gamma(x, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}} \exp\left(-(4\gamma)^{-1} \|y - x + \gamma \nabla V(x)\|^2\right) \mathrm{d}y \ . \tag{14}$$

The sequence $\{X_n^{\mathrm{M}}, \ n \in \mathbb{N}\}$ is an inhomogeneous Markov chain associated with the sequence of kernel $\{R_{\gamma_n}, \ n \geq 1\}$. Denote for $1 \leq n \leq p$,

$$Q_\gamma^{n,p} = R_{\gamma_n} \cdots R_{\gamma_p} \ , \quad Q_\gamma^n = Q_\gamma^{1,n} \ . \tag{15}$$

By convention for $n > p \geq 0$, we set $Q_\gamma^{n,p} = \mathrm{Id}$ where Id is the identity kernel. Denote for $n, p \in \mathbb{N}$ by

$$\Gamma_{n,p} = \sum_{k=n}^{p} \gamma_k \ , \quad \Gamma_n = \Gamma_{1,n} \ . \tag{16}$$

In this section we derive non-asymptotic bounds on the difference between $\mu_0 Q_\gamma^n$ and μ in total variation. Such bounds have been obtained in Dalalyan [2014] and Durmus and Moulines [2015], but the assumptions considered here are weaker. Similar to what is done in Durmus and Moulines [2015], we consider the following decomposition: for all $n \geq 0$, $p \geq 1$ and $n < p$,

$$\|\mu_0 Q_\gamma^p - \mu\|_{\mathrm{TV}} \leq \|\mu_0 Q_\gamma^n \mathbf{P}^L_{\Gamma_{n+1,p}} - \mu\|_{\mathrm{TV}} + \|\mu_0 Q_\gamma^n Q_\gamma^{n+1,p} - \mu_0 Q_\gamma^n \mathbf{P}^L_{\Gamma_{n+1,p}}\|_{\mathrm{TV}} \ , \tag{17}$$

where $(\mathbf{P}_t^L)_{t \geq 0}$ is the Markov semigroup of the Langevin SDE associated with V. The first term will be bounded using new quantitative results on the convergence of $(\mathbf{P}_t^L)_{t \geq 0}$ to π in total variation, given in Section 7. On the other hand, the second term will be bounded using the Pinsker inequality allowing to compare the total variation distance with the relative entropy combined with the Girsanov theorem; see Dalalyan [2014] and Durmus and Moulines [2015]. To complete this step, it is required to control some moments of the gradient of the potential, which is achieved by using establishing a drift conditions for R_γ with $\gamma > 0$; see Section 8.

Theorem 3. *Assume* ***H****2(V). Let* $\{\gamma_k,\ k \in \mathbb{N}^*\}$ *be a nonincreasing sequence with* $\gamma_1 \leq L_{\mathrm{c}}^{-1}$. *Then, for all* $n \geq 0$, $p \geq 1$, $n < p$, *and* $x \in \mathbb{R}^d$

$$\|\delta_x Q_\gamma^p - \mu\|_{\mathrm{TV}} \leq \mathrm{C_c} \kappa_{\mathrm{c}}^{\Gamma_{n+1,p}} \left\{ \beta_{\mathrm{c}} \theta_{\mathrm{c}}^{-1} + \lambda_{\mathrm{c}}^{\Gamma_{1,n}} \mathrm{W_c}(x) + \mathrm{c_c}(1-\lambda_{\mathrm{c}}^{\Gamma_{1,n}})/(1-\lambda_{\mathrm{c}}^{\gamma_1}) \right\} + A_{n,p}(x;\gamma) \ ,$$

where $\mathrm{C_c}, \kappa_{\mathrm{c}}, \theta_{\mathrm{c}}, \beta_{\mathrm{c}}$ *are given in Theorem 12,* λ_{c} *in (53),* $\mathrm{c_c}$ *in (57),* $\mathrm{W_c} : \mathbb{R}^d \to \mathbb{R}$ *is defined for all* $x \in \mathbb{R}^d$ *by*

$$\mathrm{W_c}(x) = \exp\left(\rho_{\mathrm{c}} \left\{ \|x - x^\star\| + 1 \right\}^{1/2} \right) \ , \tag{18}$$

$$(A_{n,p}(x;\gamma))^2 = 2^{-1} L_{\mathrm{c}}^2 \sum_{k=n}^{p-1} \left\{ (\gamma_{k+1}^3/3) L_{\mathrm{c}}^2 \left(4\rho_{\mathrm{c}}^{-1} \left\{ 1 + \log\left\{ \mathrm{W_c}(x) + \mathrm{c_c}\left(1-\lambda_{\mathrm{c}}^{\gamma_1}\right)^{-1} \right\} \right\} \right)^2 + d\gamma_{k+1}^2 \right\} \ .$$

Proof. The proof is postponed to Section 8.1. □

More precise bounds can be obtained under more stringent assumption on V. We consider the case where V is strongly convex outside some ball; see Eberle [2015].

H3 (V). *There exist* $\mathrm{R_s} \geq 1$ *and* $m_{\mathrm{s}} > 0$, *such that for all* $x, y \in \mathbb{R}^d$, $\|x - y\| \geq \mathrm{R_s}$,

$$\langle \nabla V(x) - \nabla V(y), x - y \rangle \geq 2 m_{\mathrm{s}} \|x - y\|^2 \ .$$

Note that in the case where $g = \iota_{\mathcal{K}}$, where $\mathcal{K}$ is a convex body, then by (9) and the Cauchy-Schwarz inequality, we have $\langle \nabla g^\lambda(x) - \nabla g^\lambda(y), x - y \rangle \geq \lambda^{-1}(\|x-y\|^2 - 2\{\sup_{z\in\mathcal{K}} \|z\|\}\|x-y\|)$, which implies **H3**$(V)$.

Theorem 4. *Assume* ***H****2(V) and* ***H****3(V). Let* $\{\gamma_k,\ k \in \mathbb{N}^*\}$ *be a nonincreasing sequence with* $\gamma_1 \leq 4 m_{\mathrm{s}} L_{\mathrm{c}}^{-1}$. *Then, for all* $n \geq 0$, $p \geq 1$, $n < p$, *and* $x \in \mathbb{R}^d$

$$\|\delta_x Q_\gamma^p - \mu\|_{\mathrm{TV}} \leq \mathrm{C_s} \kappa_{\mathrm{s}}^{2\Gamma_{n+1,p}} B_{n,p}(x;\gamma) + C_{n,p}(x;\gamma) \ ,$$

where $\mathrm{C_s}, \kappa_{\mathrm{s}}$ *are given in Theorem 13,* λ_{s} *in (59),* $\mathrm{c_s}$ *in (61),*

$$B_{n,p}(x;\gamma) = \left\{ 1 + \left(\frac{d}{2m_{\mathrm{s}}} + \mathrm{R_s} \right)^{1/2} + \left(\lambda_{\mathrm{s}}^{\Gamma_{1,n}} \|x - x^\star\|^2 + \mathrm{c_s} \frac{1 - \lambda_{\mathrm{s}}^{\Gamma_{1,n}}}{1 - \lambda_{\mathrm{s}}^{\gamma_1}} \right)^{1/2} \right\}$$

$$(C_{n,p}(x;\gamma))^2 = 2^{-1} L_{\mathrm{c}}^2 \sum_{k=n}^{p-1} \left\{ (\gamma_{k+1}^3/3) L_{\mathrm{c}}^2 \left\{ \|x - x^\star\|^2 + \mathrm{c_s}(1-\lambda_{\mathrm{s}}^{\gamma_1})^{-1} \right\} + d\gamma_{k+1}^2 \right\} \ .$$

Proof. The proof is postponed to Section 8.2. □

In the case of constant stepsizes $\gamma_k = \gamma$ for all $k \geq 0$, we can achieve a level of precision $\varepsilon > 0$, *i.e.* $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \varepsilon$, setting $n = 0$ and carefully choosing γ and $p \geq 1$ in the bounds given by Theorem 3 and Theorem 4. We summarise the dependence of γ and p in function of the dimension d, the precision ε and the other parameters of V in Table 1. The following result gives also the order in the stepsize γ of the asymptotic bias.

Proposition 5. *Assume* ***H****2(V) and let* $\gamma \leq L_{\mathrm{c}}$. *Then* R_γ *has a unique invariant probability measure* μ_γ, *which satisfies* $\|\mu - \mu_\gamma\|_{\mathrm{TV}} = \mathcal{O}((\gamma \log(\gamma))^{1/2})$.

Proof. The proof is postponed to 8.3. □

	d	ε	L_{c}
γ	$\mathcal{O}(d^{-4})$	$\mathcal{O}(\varepsilon^2/\log(\varepsilon^{-1}))$	$\mathcal{O}(L_{\mathrm{c}}^{-2})$
p	$\mathcal{O}(d^{7})$	$\mathcal{O}(\varepsilon^{-2}\log^2(\varepsilon^{-1}))$	$\mathcal{O}(L_{\mathrm{c}}^{2})$

Table 1: For constant stepsizes, dependence of γ and p in d, ε and parameters of V to get $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \varepsilon$ using Theorem 3

	d	ε	L_{c}	m_{s}	R_{s}
γ	$\mathcal{O}(d^{-1})$	$\mathcal{O}(\varepsilon^2/\log(\varepsilon^{-1}))$	$\mathcal{O}(L_{\mathrm{c}}^{-2})$	$\mathcal{O}(m_{\mathrm{s}})$	$\mathcal{O}(\mathrm{R}_{\mathrm{s}}^{-4})$
p	$\mathcal{O}(d\log(d))$	$\mathcal{O}(\varepsilon^{-2}\log^2(\varepsilon^{-1}))$	$\mathcal{O}(L_{\mathrm{c}}^{2})$	$\mathcal{O}(m_{\mathrm{s}}^{-2})$	$\mathcal{O}(\mathrm{R}_{\mathrm{s}}^{8})$

Table 2: For constant stepsizes, dependence of γ and p in d, ε and parameters of V to get $\|\delta_x Q_\gamma^p - \pi\|_{\mathrm{TV}} \leq \varepsilon$ using Theorem 4

4 Numerical illustrations

In this section, the behavior of the MYULA algorithm is demonstrated on a Bayesian image deconvolution model with a total-variation prior. This is a challenging high-dimensional and non-smooth model that is widely used in statistical image processing and for which the state-of-the-art MCMC algorithms are inefficient.

In image deconvolution or deblurring problems, the goal is to recover an original image $\boldsymbol{x} \in \mathbb{R}^n$ from a blurred and noisy observed image $\boldsymbol{y} \in \mathbb{R}^n$ related to $\boldsymbol{x}$ by the linear observation model $\boldsymbol{y} = H\boldsymbol{x} + \boldsymbol{w}$, where H is a linear operator representing the blur point spread function and $\boldsymbol{w}$ is a Gaussian vector with zero-mean and covariance matrix $\sigma^2 \boldsymbol{I}_n$. This inverse problem is usually ill-posed or ill-conditioned, i.e., either H does not admit an inverse or it is nearly singular, thus yielding highly noise-sensitive solutions. Bayesian image deconvolution methods address this difficulty by exploiting prior knowledge about $\boldsymbol{x}$ in order to obtain more robust estimates. One of the most widely used image prior for deconvolution problems is the improper total-variation norm prior, $\pi(\boldsymbol{x}) \propto \exp\left(-\alpha\|\nabla_d \boldsymbol{x}\|_1\right)$, where ∇_d denotes the discrete gradient operator that computes the vertical and horizontal differences between neighbour pixels. This prior encodes the fact that differences between neighbour image pixels are often very small and occasionally take large values (i.e., image gradients are nearly sparse). Based on this prior and on the linear observation model described above, the posterior distribution for $\boldsymbol{x}$ is given by

$$\pi(\boldsymbol{x}|\boldsymbol{y}) \propto \exp\left[-\|\boldsymbol{y} - H\boldsymbol{x}\|^2/2\sigma^2 - \alpha\|\nabla_d \boldsymbol{x}\|_1\right]. \tag{19}$$

Image processing methods using (19) are almost exclusively based on MAP estimates of $\boldsymbol{x}$ that can be efficiency computed using proximal optimisation algorithms [Green et al., 2015]. Here we consider the problem of computing credibility regions for $\boldsymbol{x}$, which we use to assess the confidence in the restored image. Precisely, we use MYULA to compute approximately the marginal 90% credibility interval for each image pixel, where we note that (19) is log-concave and admits the decomposition $U(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})$, with $f(\boldsymbol{x}) = -\|\boldsymbol{y} - H\boldsymbol{x}\|^2/2\sigma^2$ convex and Lipschitz differentiable, and $g(\boldsymbol{x}) = -\alpha\|\nabla_d \boldsymbol{x}\|_1$ convex and with computationally tractable proximity mapping that can be compute efficiently by using a parallel implementation of Chambolle [2004]).

Figure 1 presents an experiment with the `Boat` image, which is a standard image to assess deconvolution methods [Green et al., 2015]. Figure 1(a) and (b) show the original image $\boldsymbol{x}_0$ of size 256×256 and a blurred and noisy observation $\boldsymbol{y}$, which we produced by convoluting $\boldsymbol{x}_0$ with

a uniform blur of size 9×9 and adding white Gaussian noise to achieve a blurred signal-to-noise ratio (BSNR) of 40dB ($BRSN = 10 \log_{10}\{\mathrm{var}(H\boldsymbol{x}_0)/\sigma^2\}$). The MAP estimate of $\boldsymbol{x}$ obtained by maximising (19) is depicted in Figure 1(c). This estimate has been computed with the proximal optimisation algorithm of Afonso et al. [2011], and by using the technique of Oliveira et al. [2009] to determine the value of α. By comparing Figures 1(a) and 1(c) we observe that this image restoration process has produced a remarkably sharp image with very noticeable fine detail.

(a) (b) (c)

Figure 1: (a) Original `Boat` image (256×256 pixels), (b) Blurred image, (c) MAP estimate computed with [Afonso et al., 2011].

Moreover, Section 4 shows the magnitude of the marginal 90% credibility regions for each pixels, as measured by the distance between the 95% and 5% quantile estimates. For benchmark purpose, Section 4(a) shows the estimates obtained by using the proximal MALA Pereyra [2015]. These estimates where computed from a 10 000-sample chain generated with a thinning factor of 1 000 to reduce the algorithm's memory foot-print, and by setting $\lambda = \gamma/2$ and adjusting $\gamma = 0.004$ to achieve an acceptance rate of approximate 50%. Figures 4(b) and Figure 4(c) show respectively the approximate estimates obtained from 10 000-sample chains generated with MYULA using $\gamma = 0.02$ and a thinning factor of 100, and $\gamma = 0.2$ a thinning factor of 10 (notice that from the viewpoint of the diffusion process, the chains generated with MYULA and proximal MALA have evolved during the same "diffusion time"). Computing these estimates required approximated 35 hours for proximal MALA, 3.5 hours for MYULA with $\lambda = 0.01$, $\gamma = 0.02$ and 100-thinning, and 20 minutes for MYULA with $\lambda = 0.1$, $\gamma = 0.2$ and 10-thinning. By comparing Figures 4(a)-(c) we observe that the approximate estimates delivered by MYULA are in good agreement with the estimations obtained with proximal MALA, and with a reduction in computing time of a factor of 10 and 100.

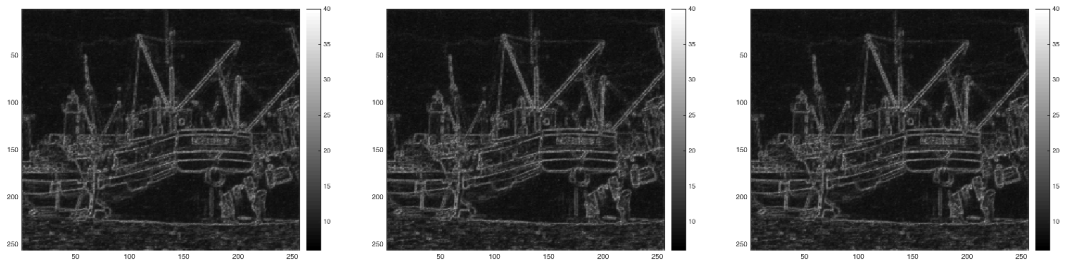
5 Proofs of Section 2

5.1 Proof of Proposition 1

We preface the proof by a Lemma.

Lemma 6. *Let* $\mathrm{h} : \mathbb{R}^d \to (-\infty, +\infty]$ *be a lower bounded, l.s.c convex function satisfying* $0 < \int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{h}(y)} \mathrm{d}y < +\infty$. *Then there exists* $x_\mathrm{h} \in \mathbb{R}^d$, $R_\mathrm{h}, \rho_\mathrm{h} > 0$ *such that for all* $x \in \mathbb{R}^d$, $x \notin \mathrm{B}(x_\mathrm{h}, R_\mathrm{h})$, $\mathrm{h}(x) - \mathrm{h}(x_\mathrm{h}) \geq \rho_\mathrm{h} \left\| x - x_\mathrm{h} \right\|$.

Proof. The proof is a simple extension of the one of [Bakry et al., 2008, Theorem 2.2.2], where h is assumed to be continuously differentiable.



(a) (b) (c)

Figure 2: (a) Pixel-wise 90% credibility intervals computed with proximal MALA (computing time 35 hours), (b) Approximate intervals estimated with MYULA using $\lambda = 0.01$ (computing time 3.5 hours), (c) Approximate intervals estimated with MYULA using $\lambda = 0.1$ (computing time 20 minutes).

We first show that h is finite on a non-empty open set of $\mathbb{R}^d$. Note since $\int_{\mathbb{R}^d} \mathrm{e}^{-\mathrm{h}(y)}\mathrm{d}y > 0$, the set $\{\mathrm{h} < \infty\}$ can not be contained in a k-dimensional hyperplane, for $k \in \{0, \cdots, d-1\}$. Then, there exists $d+1$ points $\{\mathrm{v}_i\}_{0\leq i\leq d} \subset \{\mathrm{h} < \infty\}$ such that the vectors $\{\mathrm{v}_i - \mathrm{v}_0\}_{1\leq i\leq d}$ are linearly independent. Denote by $\mathrm{co}(\mathrm{v}_0, \cdots, \mathrm{v}_d)$ the convex hull of $\{\mathrm{v}_i\}_{0\leq i\leq d}$ defined by

$$\mathrm{co}(\mathrm{v}_0, \cdots, \mathrm{v}_d) = \left\{ \sum_{i=0}^{d} \alpha_i \mathrm{v}_i \mid \sum_{i=0}^{d} \alpha_i = 1 \, , \forall i \in \{0, \cdots, d\} \, , \; \alpha_i \geq 0 \right\} .$$

Since h is convex, $\mathrm{co}(\mathrm{v}_0, \cdots, \mathrm{v}_d) \subset \{\mathrm{h} < \infty\}$ and we have

$$\sup_{y \in \mathrm{co}(\mathrm{v}_0, \cdots, \mathrm{v}_d)} |\mathrm{h}(y)| \leq M_{\mathrm{co}} = \max_{i \in \{0, \cdots, d\}} \{|\mathrm{h}(\mathrm{v}_i)|\} . \tag{20}$$

It follows from $\{\mathrm{v}_i\}_{0\leq i\leq d} \subset \{\mathrm{h} < \infty\}$ and h is lower bounded that M_{co} is finite. Finally by [Florenzano and Le Van, 2001, Lemma 1.2.1], $\mathrm{co}(\mathrm{v}_0, \cdots, \mathrm{v}_d)$ has non empty interior.

Consider now the set $\{\mathrm{h} \leq M_{\mathrm{co}}+1\}$. We prove by contradiction that it is a bounded subset of $\mathbb{R}^d$. Assume that for all $R \geq 0$, there exists $x_R \in \{\mathrm{h} \leq M_{\mathrm{co}} + 1\}$ and $x_R \notin \mathrm{B}(\mathrm{v}_0, R)$. Then since $\{\mathrm{h} \leq M_{\mathrm{co}}+1\}$ is convex, it contains the convex hull of $\{\mathrm{v}_0, \cdots, \mathrm{v}_d, x_R\}$. Since $\mathrm{co}(\mathrm{v}_0, \cdots, \mathrm{v}_d)$ has non empty interior, the volume of $\mathrm{co}(\mathrm{v}_0, \cdots, \mathrm{v}_d, x_R)$ grows at least linearly in R and the volume corresponding to $\{\mathrm{h} \leq M_{\mathrm{co}} + 1\}$ is infinite taking the limit as R goes to ∞. On the other hand, by assumption since $\{\mathrm{v}_0, \cdots, \mathrm{v}_d, x_R\} \subset \{\mathrm{h} \leq M_{\mathrm{co}} + 1\}$, we have using the Markov inequality

$$\mathrm{Vol}\left(\{\mathrm{h} \leq M_{\mathrm{co}} + 1\}\right) \leq \mathrm{e}^{M_{\mathrm{co}}+1} \int_{\{\mathrm{h}\leq M_{\mathrm{co}}+1\}} \mathrm{e}^{-\mathrm{h}(y)}\mathrm{d}y < +\infty \, ,$$

which is leads to a contradiction. Then there exists $R_{\mathrm{h}} \geq 0$, such that $\{\mathrm{h} \leq M_{\mathrm{co}}+1\} \subset \mathrm{B}(\mathrm{v}_0, R_{\mathrm{h}})$. For all $x \notin \mathrm{B}(\mathrm{v}_0, R_{\mathrm{h}})$, consider $y = R_{\mathrm{h}}(x - \mathrm{v}_0) \left\|x - \mathrm{v}_0\right\|^{-1} + \mathrm{v}_0$. Note that $y \notin \{\mathrm{h} \leq M_{\mathrm{co}} + 1\}$, so $\mathrm{h}(y) \geq M_{\mathrm{co}} + 1$. Now using the convexity of h, we have for all $x \notin \mathrm{B}(\mathrm{v}_0, R_{\mathrm{h}})$,

$$M_{\mathrm{co}} + 1 \leq \mathrm{h}(y) \leq R_{\mathrm{h}} \left\|x - \mathrm{v}_0\right\|^{-1} (\mathrm{h}(x) - \mathrm{h}(\mathrm{v}_0)) + \mathrm{h}(\mathrm{v}_0) \, .$$

Since $\mathrm{h}(\mathrm{v}_0) \leq M_{\mathrm{co}}$, we get

$$(\mathrm{h}(x) - \mathrm{h}(\mathrm{v}_0)) \geq R_{\mathrm{h}}^{-1} \left\|x - \mathrm{v}_0\right\|$$

and the proof is concluded setting $x_{\mathrm{h}} = \mathrm{v}_0$. □

Proof of Proposition 1. By (5), $U \geq U^\lambda$ and therefore $0 < \int_{\mathbb{R}^d} \mathrm{e}^{-U(y)} \mathrm{d}y < \int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(y)} \mathrm{d}y$. We now prove e^{-g^λ} is integrable with respect to the Lebesgue measure, which implies $y \mapsto \mathrm{e}^{-U^\lambda(y)}$ is integrable as well since f is assumed to be lower bounded. By **H**1 and Lemma 6, there exist $\rho_g > 0$, $x_g \in \mathbb{R}^d$ and $M_1 \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$, $g(x) - g(x_g) \geq M_1 + \rho_g \|x - x_g\|$. Thus, for all $x \in \mathbb{R}^d$, we have

$$\begin{aligned} g^\lambda(x) - g(x_g) &\geq M_1 + \rho_g \left\|\mathrm{prox}_g^\lambda(x) - x_g\right\| + (2\lambda)^{-1} \left\|x - \mathrm{prox}_g^\lambda(x)\right\|^2 \\ &\geq M_1 + \inf_{y \in \mathbb{R}^d} \{\rho_g \|y - x_g\| + (2\lambda)^{-1} \|x - y\|^2\} \geq M_1 + \mathrm{h}^\lambda(x) \;, \end{aligned} \tag{21}$$

where $\mathrm{h}^\lambda(x)$ is the λ-Moreau Yosida envelope of $\mathrm{h}(x) = \rho_g \|x - x_g\|$. By [Parikh and Boyd, 2013, Section 6.5.1], the proximal operator associated with the norm is the block soft thresholding given for all $\lambda > 0$ and $x \in \mathbb{R}^d \setminus \{0\}$ by $\mathrm{prox}_{\mathrm{h}}^\lambda(x) = \max(0, 1 - \lambda/\|x\|)x$ and $\mathrm{prox}_{\mathrm{h}}^\lambda(0) = 0$. Therefore, it follows that there exists M_2 such that for all $x \in \mathbb{R}^d$,

$$\mathrm{h}^\lambda(y) \geq \rho_g \|y - x_g\| + M_2 \;.$$

Combining this inequality with (21) concludes the proof. □

5.2 Proof of Proposition 2

(a) For ease of notation we also denote by π^λ the density of π^λ with respect to the Lebesgue measure. Since π has also a density with respect to the Lebesgue measure and $U^\lambda(x) \leq U(x)$ for all $x \in \mathbb{R}^d$, we have for all $\lambda > 0$

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} = \int_{\mathbb{R}^d} \left|\pi^\lambda(x) - \pi(x)\right| \mathrm{d}x \leq 2A_\lambda \;, \tag{22}$$

where

$$A_\lambda = \int_{\mathbb{R}^d} \{1 - \mathrm{e}^{g^\lambda(x) - g(x)}\} \pi^\lambda(x) \mathrm{d}x = 1 - \left\{\int_{\mathbb{R}^d} \mathrm{e}^{-U^\lambda(x)} \mathrm{d}x\right\}^{-1} \int_{\mathbb{R}^d} \mathrm{e}^{-U(x)} \mathrm{d}x \;.$$

By (7), for all $x \in \mathbb{R}^d$, we get $\lim_{\lambda \downarrow 0} \uparrow U^\lambda(x) = U(x)$. We conclude by applying the monotone convergence theorem.

(b) Using that for all $x \in \mathbb{R}^d$, $g^\lambda(x) \leq g(x)$ and $1 - \mathrm{e}^{-u} \leq u$ for all $u \geq 0$, (22) shows that

$$\|\pi^\lambda - \pi\|_{\mathrm{TV}} \leq 2 \int_{\mathbb{R}^d} \{g(x) - g^\lambda(x)\} \pi^\lambda(x) \mathrm{d}x \;.$$

Next, we show that $\sup_{x \in \mathbb{R}^d} \{g(x) - g^\lambda(x)\} \leq \lambda \|g\|_{\mathrm{Lip}}^2 / 2$, which will conclude the proof. Using that g is Lipschitz, we have by (5), for all $x \in \mathbb{R}^d$

$$\begin{aligned} g(x) - g^\lambda(x) &= g(x) - \inf_{y \in \mathbb{R}^d} \left\{g(y) + (2\lambda)^{-1} \|x - y\|^2\right\} \\ &= \sup_{y \in \mathbb{R}^d} \left\{g(x) - g(y) - (2\lambda)^{-1} \|x - y\|^2\right\} \\ &\leq \sup_{y \in \mathbb{R}^d} \left\{\|g\|_{\mathrm{Lip}} \|x - y\| - (2\lambda)^{-1} \|x - y\|^2\right\} \leq \lambda \|g\|_{\mathrm{Lip}}^2 / 2 \;, \end{aligned}$$

where we have used that the maximum of $u \mapsto au - bu^2$, for $a, b \geq 0$, is given by $a^2/(4b)$.

(c) Consider the intrinsic volumes $\{\mathscr{V}_i(\mathcal{K})\}_{0\le i\le d}$ of $\mathcal{K}$, which can be defined by Steiner's formula, which states that for all $M \ge 0$,

$$\mathrm{Vol}(\mathcal{K} + \mathrm{B}(0, M)) = \sum_{i=0}^{d} M^{d-i} \boldsymbol{\pi}^{(d-i)/2} \boldsymbol{\Gamma}^{-1}(1 + (d-i)/2) \mathscr{V}_i(\mathcal{K}) \,, \tag{23}$$

where $\boldsymbol{\Gamma} : \mathbb{R}_+^* \to \mathbb{R}_+^*$ is the Gamma function. We refer to [Schneider, 2013, Chapter 4.2] for this result and an introduction to this topic. By definition of the projection onto a closed convex subset of $\mathbb{R}^d$, g^λ is given for all $x \in \mathbb{R}^d$ by $g^\lambda(x) = (2\lambda)^{-1} \left\| x - \mathrm{proj}_{\mathcal{K}}(x) \right\|^2$. Then by a direct calculation, we get

$$\begin{aligned} \|\pi^\lambda - \pi\|_{\mathrm{TV}} &= \int_{\mathbb{R}^d} \left| \pi(x) - \pi^\lambda(x) \right| \mathrm{d}x = 2 \left(1 + \left\{ \int_{\mathcal{K}^c} \mathrm{e}^{-U^\lambda(x)} \mathrm{d}x \right\}^{-1} \int_{\mathcal{K}} \mathrm{e}^{-f(x)} \mathrm{d}x \right)^{-1} \\ &\le 2 \left(1 + \exp\left(\min_{\mathcal{K}^c}(f) - \max_{\mathcal{K}}(f) \right) \left\{ \mathrm{Vol}(\mathcal{K}) \Big/ \int_{\mathcal{K}^c} \mathrm{e}^{-(2\lambda)^{-1} \|x - \mathrm{proj}_{\mathcal{K}}(x)\|^2} \mathrm{d}x \right\} \right)^{-1} . \end{aligned} \tag{24}$$

In addition using [Kampf, 2009, Proposition 3], we get

$$\int_{\mathcal{K}^c} \mathrm{e}^{-(2\lambda)^{-1} \|x - \mathrm{proj}_{\mathcal{K}}(x)\|^2} \mathrm{d}x = \int_{\mathbb{R}_+} \mathrm{Vol}\left(\mathcal{K} + \mathrm{B}(0, t) \right) \mathrm{e}^{-t^2/(2\lambda)} \mathrm{d}t \,. \tag{25}$$

Combining (24)-(25) and Steiner's formula (23) implies that (12) holds with

$$\mathsf{D}(\mathcal{K}, \lambda) = \exp\left(\min_{\mathcal{K}^c}(f) - \max_{\mathcal{K}}(f) \right) \mathrm{Vol}(\mathcal{K}) \left\{ \sum_{i=0}^{d-1} (\lambda/2\pi)^{(d-i)/2} \mathscr{V}_i(\mathcal{K}) \right\}^{-1} .$$

6 Proof of Section 3

7 Quantitative convergence bounds in total variation for diffusions

In this part, we are interested in quantitative convergence results in total variation norm for d-dimensional SDEs of the form

$$\mathrm{d}\mathbf{X}_t = b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}B_t^d \,, \tag{26}$$

started at $\mathbf{X}_0$, where $(B_t^d)_{t\ge 0}$ is a d-dimensional standard Brownian motion and $b : \mathbb{R}^d \to \mathbb{R}^d$ satisfies the following assumptions.

G 1 (b)**.** *b is Lipschitz and for all $x, y \in \mathbb{R}^d$, $\langle b(x) - b(y), x - y \rangle \le 0$.*

Under **G**1(b), [Ikeda and Watanabe, 1989, Theorems 2.4-3.1-6.1, Chapter IV] imply that there exists a unique solution $(\mathbf{X}_t)_{t\ge 0}$ to (26) for all initial distribution ξ_0 on $\mathbb{R}^d$, which is strongly Markovian. Denote by $(\mathbf{P}_t)_{t\ge 0}$ the transition semigroup associated with (26). To derive explicit bound for $\|\mathbf{P}_t(x,\cdot) - \mathbf{P}_t(y,\cdot)\|_{\mathrm{TV}}$, we use the coupling by reflection, introduced in Lindvall and

Rogers [1986]. This coupling is defined as (see [Chen and Li, 1989, Example 3.7]) the unique strong Markovian process $(\mathbf{X}_t, \mathbf{Y}_t)_{t\geq 0}$ on $\mathbb{R}^{2d}$, solving the SDE:

$$\begin{cases} \mathrm{d}\mathbf{X}_t &= b(\mathbf{X}_t)\mathrm{d}t + \mathrm{d}B_t^d \\ \mathrm{d}\mathbf{Y}_t &= b(\mathbf{Y}_t)\mathrm{d}t + (\mathrm{Id} - 2e_t e_t^T)\mathrm{d}B_t^d \end{cases}, \qquad \text{where } e_t = e(\mathbf{X}_t - \mathbf{Y}_t) \tag{27}$$

with $e(z) = z/\|z\|$ for $z \neq 0$ and $e(0) = 0$ otherwise. Define the coupling time

$$\tau_c = \inf\{s \geq 0 \mid \mathbf{X}_s \neq \mathbf{Y}_s\} . \tag{28}$$

By construction $\mathbf{X}_t = \mathbf{Y}_t$ for $t \geq \tau_c$. We denote in the sequel by $\tilde{\mathbb{P}}_{(x,y)}$ and $\tilde{\mathbb{E}}_{(x,y)}$ the probability and the expectation associated with the SDE (27) started at $(x, y) \in \mathbb{R}^{2d}$ on the canonical space of continuous function from $\mathbb{R}_+$ to $\mathbb{R}^{2d}$. We denote by $(\widetilde{\mathcal{F}}_t)_{t\geq 0}$ the canonical filtration. Since $\bar{B}_t^d = \int_0^t (\mathrm{Id} - 2\mathbb{1}_{\{s<\tau_c\}} e_s e_s^T)\mathrm{d}B_s^d$ is a d-dimensional Brownian motion, the marginal processes $(\mathbf{X}_t)_{t\geq 0}$ and $(\mathbf{Y}_t)_{t\geq 0}$ are weak solutions to (26) started at x and y respectively. Coupling by reflection was introduced in Lindvall and Rogers [1986] to show convergence in total variation norm for solution of SDE, and recently used by Eberle [2015] to obtain exponential convergence in the Wasserstein distance of order 1. The result in Lindvall and Rogers [1986] are derived under less stringent conditions than **G**1(b), but do not provide quantitative estimates.

Proposition 7 ([Lindvall and Rogers, 1986, Example 5]). *Assume* ***G****1(b) and let $(\mathbf{X}_t, \mathbf{Y}_t)_{t\geq 0}$ be the solution of (27). Then for all $t \geq 0$ and $x, y \in \mathbb{R}^d$, we have*

$$\tilde{\mathbb{P}}_{(x,y)}\left(\tau_c \geq t\right) = \tilde{\mathbb{P}}_{(x,y)}\left(\mathbf{X}_t \neq \mathbf{Y}_t\right) \leq 2\left(\mathbf{\Phi}\left\{\left(2t^{1/2}\right)^{-1}\|x - y\|\right\} - 1/2\right) .$$

Proof. For $t < \tau_c$, $\mathbf{X}_t - \mathbf{Y}_t$ is the solution of the SDE

$$\mathbf{X}_t - \mathbf{Y}_t = \{b(\mathbf{X}_t) - b(\mathbf{Y}_t)\}\,\mathrm{d}t + 2e_t\mathrm{d}B_t^1 ,$$

where $(B_t^1)_{t\geq 0}$ is the one-dimensional Brownian motion given by $B_t^1 = \int_0^t \mathbb{1}_{\{s<\tau_c\}} e_s^T \mathrm{d}B_s^d$. Using the Itô's formula and **G**1, we have for all $t < \tau_c$,

$$\|\mathbf{X}_t - \mathbf{Y}_t\| = \|x - y\| + \int_0^t \langle b(\mathbf{X}_s) - b(\mathbf{Y}_s), e_s\rangle\,\mathrm{d}s + 2B_t^1 \leq \|x - y\| + 2B_t^1 . \tag{29}$$

Therefore, for all $x, y \in \mathbb{R}^d$ and $t \geq 0$, we get

$$\begin{aligned} \tilde{\mathbb{P}}_{(x,y)}\left(\tau_c > t\right) &\leq \tilde{\mathbb{P}}_{(x,y)}\left(\min_{0\leq s\leq t} B_s^1 \geq \|x - y\|/2\right) \\ &= \tilde{\mathbb{P}}_{(x,y)}\left(\max_{0\leq s\leq t} B_s^1 \leq \|x - y\|/2\right) = \tilde{\mathbb{P}}_{(x,y)}\left(|B_t^1| \leq \|x - y\|/2\right) . \end{aligned}$$

where we have used the reflection principle in the last identity. □

Define the set

$$\Delta_R = \{x, y \in \mathbb{R}^d \mid \|x - y\| \leq R\} \tag{30}$$

for $R > 0$ and the function $\omega : (0, 1) \times \mathbb{R}_+^* \to \mathbb{R}_+$ by

$$\omega(\epsilon, R) = R^2 / \left\{2\mathbf{\Phi}^{-1}(1 - \epsilon/2)\right\}^2 . \tag{31}$$

Proposition 7 and Lindvall's inequality give that, for all $\epsilon \in (0,1)$,

$$\sup_{(x,y)\in\Delta_R} \|\mathbf{P}_t(x,\cdot) - \mathbf{P}_t(y,\cdot)\|_{\mathrm{TV}} \leq (1-\epsilon) \ , \ \text{for any } t \geq \omega(\epsilon, R) \ , \tag{32}$$

To obtain quantitative exponential bounds in total variation for any $x, y \in \mathbb{R}^d$, it is required to control some exponential moments of the return times to Δ_R. This is first achieved by using a drift condition for the generator $\mathscr{A}$ associated with the SDE (26) defined for all $f \in C^2(\mathbb{R}^d)$ by

$$\mathscr{A} f = \langle b, \nabla f \rangle + (1/2)\Delta f \ . \tag{33}$$

Consider the following assumption:

G 2 $(\mathsf{W}, \theta, \beta)$. *(i) There exist* $\mathsf{W} \in C^2(\mathbb{R}^d)$, $\mathsf{W} \geq 1$, *and* $\theta > 0$, $\beta \geq 0$ *such that*

$$\mathscr{A}\mathsf{W} \leq -\theta \mathsf{W} + \beta \ . \tag{34}$$

(ii) There exists $\delta > 0$ *and* $\mathrm{R} > 0$ *such that* $\Theta_\delta \subset \Delta_{\mathrm{R}}$ *where*

$$\Theta = \{(x,y) \in \mathbb{R}^{2d} \mid \mathsf{W}(x) + \mathsf{W}(y) \leq 2\theta^{-1}\beta + \delta\} \ . \tag{35}$$

For $t > 0$, and G a closed subset of $\mathbb{R}^{2d}$, let $\mathrm{T}_1^{G,t}$ be the first return time to G delayed by t defined by

$$\mathrm{T}_1^{G,t} = \inf\{s > t \mid (\mathbf{X}_s, \mathbf{Y}_s) \in G\} \ . \tag{36}$$

For $j \geq 2$; define recursively the j-th return times to G delayed by t by

$$\mathrm{T}_j^{G,t} = \inf\left\{s \geq \mathrm{T}_{j-1}^{G,t} + t \mid (\mathbf{X}_s, \mathbf{Y}_s) \in G\right\} = \mathrm{T}_{j-1}^{G,t} + \mathrm{T}_1^{G,t} \circ \mathrm{S}_{\mathrm{T}_{j-1}^{G,t}} \ , \tag{37}$$

where S is the shift operator on the canonical space. By [Ethier and Kurtz, 1986, Proposition 1.5 Chapter 2], the sequence $(\mathrm{T}_j^{G,t})_{j\geq 1}$ is a sequence of stopping time with respect to $(\widetilde{\mathcal{F}}_t)_{t\geq 0}$. For all $j \geq 1$ and $\epsilon \in (0,1)$, set

$$\mathrm{T}_j = \mathrm{T}_j^{\Theta, \omega(\epsilon,\mathrm{R})} \ , \tag{38}$$

where δ, R are given in $\mathbf{G}2(\mathsf{W}, \theta, \beta)$, ω in (31) and Θ in (35).

Proposition 8. *Assume* $\boldsymbol{G}1(b)$ *and* $\boldsymbol{G}2(\mathsf{W}, \theta, \beta)$. *For all* $x, y \in \mathbb{R}^d$, $\epsilon \in (0,1)$ *and* $j \geq 1$, *we have*

$$\tilde{\mathbb{E}}_{(x,y)}\left[\mathrm{e}^{\tilde{\theta}\mathrm{T}_j}\right] \leq \mathrm{K}^{j-1}\left\{(1/2)(\mathsf{W}(x) + \mathsf{W}(y)) + \mathrm{e}^{\tilde{\theta}\omega(\epsilon,\mathrm{R})}\tilde{\theta}^{-1}\beta\right\} \ ,$$

where

$$\tilde{\theta} = \theta^2\delta(2\beta + \theta\delta)^{-1} \ , \ \mathrm{K} = \tilde{\theta}^{-1}\beta\left(1 + \mathrm{e}^{\tilde{\theta}\omega(\epsilon,\mathrm{R})}\right) + \delta/2 \ . \tag{39}$$

Proof. Note that for all $x, y \in \mathbb{R}^d$,

$$\mathscr{A}\mathsf{W}(x) + \mathscr{A}\mathsf{W}(y) \leq -\tilde{\theta}(\mathsf{W}(x) + \mathsf{W}(y)) + 2\beta \mathbb{1}_\Theta(x,y) \tag{40}$$

Then by the Dynkin formula (see e.g. [Meyn and Tweedie, 1993, Eq. (8)]) the process

$$t \mapsto (1/2)\exp\left(\tilde{\theta}\left(\mathrm{T}_1 \wedge t\right)\right)\left\{\mathsf{W}\left(\mathbf{X}_{\mathrm{T}_1\wedge t}\right) + \mathsf{W}\left(\mathbf{Y}_{\mathrm{T}_1\wedge t}\right)\right\}$$

is a positive supermartingale. Using the optional stopping theorem and the Markov property, we have

$$\tilde{\mathbb{E}}_{(x,y)}\left[\mathrm{e}^{\tilde{\theta}\mathrm{T}_1}\right] \leq (1/2)(\mathsf{W}(x) + \mathsf{W}(y)) + \mathrm{e}^{\tilde{\theta}\omega(\epsilon,\mathrm{R})}\tilde{\theta}^{-1}\beta \ .$$

The result then follows from this inequality and the strong Markov property. □

Theorem 9. *Assume* ***G****1*(b) *and* ***G****2*$(\mathsf{W}, \theta, \beta)$. *Then for all* $\epsilon \in (0,1)$, $t \geq 0$ *and* $x, y \in \mathbb{R}^d$,

$$\|\mathbf{P}_t(x,\cdot) - \mathbf{P}_t(y,\cdot)\|_{\mathrm{TV}} \leq ((1-\epsilon)^{-2} + (1/2)\{\mathsf{W}(x) + \mathsf{W}(y)\})\kappa^t \ ,$$

where ω *are defined in* (31), $\tilde{\theta}, \mathrm{K}$ *in* (39) *and*

$$\log(\kappa) = \tilde{\theta}\log(1-\epsilon)(\log(\mathrm{K}) - \log(1-\epsilon))^{-1} \ .$$

Proof. Let $x, y \in \mathbb{R}^d$ and $t \geq 0$. For all $\mathrm{m} \geq 1$ and $\epsilon \in (0,1)$,

$$\tilde{\mathbb{P}}_{(x,y)}(\tau_c > t) \leq \tilde{\mathbb{P}}_{(x,y)}(\tau_c > t, \mathrm{T_m} \leq t) + \tilde{\mathbb{P}}_{(x,y)}(\mathrm{T_m} > t) \ , \tag{41}$$

where $\mathrm{T_m}$ is defined in (38). We now bound the two term in the right hand side of this equation. For the first term, since $\Theta \subset \Delta_{\mathrm{R}}$, by (32), we have conditioning successively on $\widetilde{\mathcal{F}}_{\mathrm{T}_j}$, for $j = \mathrm{m}, \ldots, 1$, and using the strong Markov property,

$$\tilde{\mathbb{P}}_{(x,y)}(\tau_c > t, \mathrm{T_m} \leq t) \leq (1-\epsilon)^{\mathrm{m}} \ . \tag{42}$$

For the second term, using Proposition 8 and the Markov inequality, we get

$$\tilde{\mathbb{P}}_{(x,y)}(\mathrm{T_m} > t) \leq \mathrm{e}^{-\tilde{\theta}t}\mathrm{K}^{\mathrm{m}-1}\left\{(1/2)(\mathsf{W}(x) + \mathsf{W}(y)) + \mathrm{e}^{\tilde{\theta}\omega(\epsilon,\mathrm{R})}\tilde{\theta}^{-1}\beta\right\} \tag{43}$$

Combining (42)-(43) in (41) and taking $\mathrm{m} = \left\lfloor \tilde{\theta}t/(\log(\mathrm{K}) - \log(1-\epsilon)) \right\rfloor$ concludes the proof. □

More precise bounds can be obtained under more stringent assumption on the drift b. We consider the case where b is strongly convex outside some ball; see Eberle [2015].

G3 (b). *There exist* $\bar{\mathrm{R}}_{\mathrm{s}} \geq 1$ *and* $\bar{m}_{\mathrm{s}} > 0$, *such that for all* $x, y \in \mathbb{R}^d$, $\|x - y\| \geq \bar{\mathrm{R}}_{\mathrm{s}}$,

$$\langle b(x) - b(y), x - y\rangle \leq -\bar{m}_{\mathrm{s}}\|x - y\|^2 \ .$$

For all $j \geq 1$ and $\epsilon \in (0,1)$, set $\mathrm{T}_j = \mathrm{T}_j^{\Delta_{\bar{\mathrm{R}}_{\mathrm{s}}},\omega(\epsilon,\bar{\mathrm{R}}_{\mathrm{s}})}$ where $\mathrm{T}_j^{\Delta_{\bar{\mathrm{R}}_{\mathrm{s}}},\omega(\epsilon,\bar{\mathrm{R}}_{\mathrm{s}})}$ is defined in (36)-(37).

Proposition 10. *Assume* ***G****1*(b) *and* ***G****3*(b).

a) For all $x, y \in \mathbb{R}^d$ *and* $\epsilon \in (0,1)$

$$\tilde{\mathbb{E}}_{(x,y)}\left[\exp\left((\bar{m}_{\mathrm{s}}/2)(\tau_c \wedge \mathrm{T}_1)\right)\right] \leq 1 + \|x - y\| + (1 + \bar{\mathrm{R}}_{\mathrm{s}})\mathrm{e}^{\bar{m}_{\mathrm{s}}\omega(\epsilon,\bar{\mathrm{R}}_{\mathrm{s}})/2} \ .$$

b) For all $x, y \in \mathbb{R}^d$, $\epsilon \in (0,1)$ *and* $j \geq 1$

$$\tilde{\mathbb{E}}_{(x,y)}\left[\exp\left((\bar{m}_{\mathrm{s}}/2)(\tau_c \wedge \mathrm{T}_j)\right)\right] \leq \mathrm{D}^{j-1}\left\{1 + \|x - y\| + (1 + \bar{\mathrm{R}}_{\mathrm{s}})\mathrm{e}^{\bar{m}_{\mathrm{s}}\omega(\epsilon,\bar{\mathrm{R}}_{\mathrm{s}})/2}\right\} \ ,$$

where D *is defined in* (47).

Proof. a) Consider the sequence of increasing stopping time $\tau_k = \inf\{t > 0 \mid k^{-1} \leq \|\mathbf{X}_t - \mathbf{Y}_t\| \leq k\}$, for $k \geq 1$ and set $\eta_k = \tau_k \wedge \mathrm{T}_1$. We derive a bound on $\tilde{\mathbb{E}}_{(x,y)}[\exp\{(\bar{m}_{\mathrm{s}}/2)\eta_k\}]$ independent on k. Since $\lim_{k\to+\infty} \uparrow \tau_k = \tau_c$ almost surely, the monotone convergence theorem implies that the same bound holds for $\tilde{\mathbb{E}}_{(x,y)}[\exp\{(\bar{m}_{\mathrm{s}}/2)(\tau_c \wedge \mathrm{T}_1)\}]$. Set now $\mathrm{W}_{\mathrm{s}}(x,y) = 1 + \|x - y\|$. Since $\mathrm{W}_{\mathrm{s}} \geq 1$ and

$\tau_c < \infty$ a.s by Proposition 7, it suffices to give a bound on $\tilde{\mathbb{E}}_{(x,y)}[\exp\{(\bar{m}_s/2)\eta_k\}\mathrm{W}_s(\mathbf{X}_{\eta_k}, \mathbf{Y}_{\eta_k})]$. By Itô's formula, we have for all $v, t \le \tau_c$, $v \le t$

$$\mathrm{e}^{\bar{m}_s t/2}\mathrm{W}_s(\mathbf{X}_t, \mathbf{Y}_t) = \mathrm{e}^{\bar{m}_s v/2}\mathrm{W}_s(\mathbf{X}_s, \mathbf{Y}_s) + (\bar{m}_s/2)\int_v^t \mathrm{e}^{\bar{m}_s u/2}\mathrm{W}_s(\mathbf{X}_u, \mathbf{Y}_u)\mathrm{d}u \\ + \int_v^t \mathrm{e}^{\bar{m}_s u/2}\left\langle b(\mathbf{X}_u) - b(\mathbf{Y}_u), e_u\right\rangle \mathrm{d}u + 2\int_v^t \mathrm{e}^{\bar{m}_s u/2}\mathrm{d}B_u^1 \,. \quad (44)$$

By (44) and **G**3(b), we have for all $k \ge 1$ and $v, t \ge 0$, $\omega(\epsilon, \bar{\mathrm{R}}_s) \le v \le t$

$$\mathrm{e}^{(\bar{m}_s/2)(\eta_k \wedge t)}\mathrm{W}_s(\mathbf{X}_{\eta_k\wedge t}, \mathbf{Y}_{\eta_k\wedge t}) \le \mathrm{e}^{(\bar{m}_s/2)(\eta_k\wedge v)}\mathrm{W}_s(\mathbf{X}_{\eta_k\wedge v}, \mathbf{Y}_{\eta_k\wedge v}) + 2\int_{\eta_k\wedge v}^{\eta_k\wedge t} \mathrm{e}^{\bar{m}_s u/2}\mathrm{d}B_u^1 \,.$$

So the process

$$\left\{\exp\left((\bar{m}_s/2)(\eta_k \wedge t)\right)\mathrm{W}_s(\mathbf{X}_{\eta_k\wedge t}, \mathbf{Y}_{\eta_k\wedge t})\right\}_{t\ge\omega(\epsilon,\bar{\mathrm{R}}_s)} \,,$$

is a positive supermartingale and by the optional stopping theorem, we get

$$\tilde{\mathbb{E}}_{(x,y)}\left[\mathrm{e}^{(\bar{m}_s/2)\eta_k}\mathrm{W}_s(\mathbf{X}_{\eta_k}, \mathbf{Y}_{\eta_k})\right] \le \tilde{\mathbb{E}}_{(x,y)}\left[\mathrm{e}^{(\bar{m}_s/2)(\tau_k\wedge\omega(\epsilon,\bar{\mathrm{R}}_s))}\mathrm{W}_s(\mathbf{X}_{\tau_k\wedge\omega(\epsilon,\bar{\mathrm{R}}_s)}, \mathbf{Y}_{\tau_k\wedge\omega(\epsilon,\bar{\mathrm{R}}_s)})\right] \,, \quad (45)$$

where we used that $\eta_k \wedge \omega(\epsilon, \bar{\mathrm{R}}_s) = \tau_k \wedge \omega(\epsilon, \bar{\mathrm{R}}_s)$. By (44), **G**1($b$) and **G**3($b$), we have

$$\tilde{\mathbb{E}}_{(x,y)}\left[\mathrm{e}^{(\bar{m}_s/2)(\tau_k\wedge\omega(\epsilon,\bar{\mathrm{R}}_s))}\mathrm{W}_s(\mathbf{X}_{\tau_k\wedge\omega(\epsilon,\bar{\mathrm{R}}_s)}, \mathbf{Y}_{\tau_k\wedge\omega(\epsilon,\bar{\mathrm{R}}_s)})\right] \le \mathrm{W}_s(x, y) + (1 + \bar{\mathrm{R}}_s)\mathrm{e}^{\bar{m}_s\omega(\epsilon,\bar{\mathrm{R}}_s)/2} \,,$$

and (45) becomes

$$\tilde{\mathbb{E}}_{(x,y)}\left[\mathrm{e}^{(\bar{m}_s/2)\eta_k}\mathrm{W}_s(\mathbf{X}_{\eta_k}, \mathbf{Y}_{\eta_k})\right] \le \mathrm{W}_s(x, y) + (1 + \bar{\mathrm{R}}_s)\mathrm{e}^{\bar{m}_s\omega(\epsilon,\bar{\mathrm{R}}_s)/2} \,,$$

which concludes the proof of the first point.

b) The proof is by induction. For $j = 1$, it is the first point. Now let $j \ge 2$. Since on the event $\{\tau_c > \mathrm{T}_{j-1}\}$, we have

$$\tau_c \wedge \mathrm{T}_j = \mathrm{T}_{j-1} + (\tau_c \wedge \mathrm{T}_1) \circ \mathrm{S}_{\mathrm{T}_{j-1}} \,,$$

where S is the shift operator, we have conditioning on $\widetilde{\mathcal{F}}_{\mathrm{T}_{j-1}}$, using the strong Markov property, Proposition 7 and the first part,

$$\tilde{\mathbb{E}}_{(x,y)}\left[\mathbb{1}_{\tau_c>\mathrm{T}_{j-1}}\exp\left((\bar{m}_s/2)\left(\tau_c \wedge \mathrm{T}_j\right)\right)\right] \le \mathrm{D}\,\tilde{\mathbb{E}}_{(x,y)}\left[\mathbb{1}_{\tau_c>\mathrm{T}_{j-1}}\exp\left((\bar{m}_s/2)\mathrm{T}_{j-1}\right)\right] \,,$$

Then the proof follows since $\mathrm{D} \ge 1$.

□

Theorem 11. *Assume* ***G****1*(b) *and* ***G****3*(b). *Then for all* $\epsilon \in (0, 1)$, $t \ge 0$ *and* $x, y \in \mathbb{R}^d$,

$$\|\mathbf{P}_t(x, \cdot) - \mathbf{P}_t(y, \cdot)\|_{\mathrm{TV}} \le \left\{(1 - \epsilon)^{-2} + 1 + \|x - y\|\right\}\kappa_s^t \,,$$

where ω *is defined in* (31) *and*

$$\log(\kappa_s) = (\bar{m}_s/2)\log(1 - \epsilon)(\log(\mathrm{D}) - \log(1 - \epsilon))^{-1} \,, \quad (46)$$

$$\mathrm{D} = (1 + \mathrm{e}^{\bar{m}_s\omega(\epsilon,\bar{\mathrm{R}}_s)/2})(1 + \bar{\mathrm{R}}_s) \,. \quad (47)$$

Proof. The proof follows the same line as the proof of Theorem 9. Let $x, y \in \mathbb{R}^d$ and $t \geq 0$. For all $\mathrm{m} \geq 1$ and $\epsilon \in (0,1)$,

$$\tilde{\mathbb{P}}_{(x,y)}\left(\tau_c > t\right) \leq \tilde{\mathbb{P}}_{(x,y)}\left(\tau_c > t, \mathrm{T_m} \leq t\right) + \tilde{\mathbb{P}}_{(x,y)}\left(\mathrm{T_m} \wedge \tau_c > t\right) \ . \tag{48}$$

For the first term, by (32) we have conditioning successively on $\widetilde{\mathcal{F}}_{\mathrm{T}_j}$, for $j = \mathrm{m}, \cdots, 1$, and using the strong Markov property,

$$\tilde{\mathbb{P}}_{(x,y)}\left(\tau_c > t, \mathrm{T_m} \leq t\right) \leq (1-\epsilon)^{\mathrm{m}} \ . \tag{49}$$

For the second term, using Proposition 10-b) and the Markov inequality, we get

$$\tilde{\mathbb{P}}_{(x,y)}\left(\mathrm{T_m} \wedge \tau_c > t\right) \leq \mathrm{e}^{-(\bar{m}_{\mathrm{s}} t/2)} \mathrm{D}^{\mathrm{m}-1} \left\{1 + \|x-y\| + (1+\bar{\mathrm{R}}_{\mathrm{s}}) \mathrm{e}^{\bar{m}_{\mathrm{s}} \omega(\epsilon, \bar{\mathrm{R}}_{\mathrm{s}})/2}\right\} \ . \tag{50}$$

Combining (49)-(50) in (48) and taking $\mathrm{m} = \left\lfloor (\bar{m}_{\mathrm{s}} t/2) / (\log(\mathrm{D}) - \log(1-\epsilon)) \right\rfloor$ concludes the proof. □

Application to the Langevin SDE

Recall that $(\mathbf{P}_t^L)_{t\geq 0}$ is the Markov semigroup of the Langevin equation associated with and let $\mathscr{A}^L$ be the corresponding generator. Since $(\mathbf{P}_t^L)_{t\geq 0}$ is reversible with respect to μ, we deduce from Theorem 9 quantitative bounds for the exponential convergence of $(\mathbf{P}_t^L)_{t\geq 0}$ to μ in total variation noting that if $(\mathbf{X}_t^L)_{t\geq 0}$ is a solution of (2), then $(\mathbf{X}_{t/2}^L)_{t\geq 0}$ is a solution of the rescaled Langevin diffusion:

$$\mathrm{d}\tilde{\mathbf{X}}_t^L = -(1/2)\nabla U(\tilde{\mathbf{X}}_t^L)\mathrm{d}t + \mathrm{d}B_t^d \ .$$

Theorem 12. *Assume **H**2(V). Then for all $t \geq 0$, $x, y \in \mathbb{R}^d$, we have*

$$\|\mathbf{P}_t^L(x,\cdot) - \mathbf{P}_t^L(y,\cdot)\|_{\mathrm{TV}} \leq \mathrm{C_c}\kappa_{\mathrm{c}}^t \left\{\mathrm{W_c}(x) + \mathrm{W_c}(y)\right\} \ , \tag{51}$$

$$\|\mathbf{P}_t^L(x,\cdot) - \mu\|_{\mathrm{TV}} \leq \mathrm{C_c}\kappa_{\mathrm{c}}^t \left\{\mathrm{W_c}(x) + \beta_{\mathrm{c}}\theta_{\mathrm{c}}^{-1}\right\} \tag{52}$$

where R, ρ is given in (13), $\mathrm{W_c}$ in (18),

$$\log(\kappa_{\mathrm{c}}) = -2\log(2)\theta_{\mathrm{c}} \left(\log\left(\beta_{\mathrm{c}}\left\{2 + 2\theta_{\mathrm{c}}^{-1}\mathrm{e}^{2\theta_{\mathrm{c}}^{-1}\omega_{\mathrm{c}}}\right\}\right) + \log(2)\right)^{-1} \ , \ \mathrm{C_c} = 5/4 \ ,$$

$$\beta_{\mathrm{c}} = (\rho/4)(\rho/4 + d + \sup_{\{y \in \mathrm{B}(x^\star, \mathrm{a_c})\}}\{\|\nabla U(y)\|\}) \times \max\left\{1, (\mathrm{a_c}^2+1)^{-1/2}\exp(\rho(\mathrm{a_c}^2+1)^{1/2}/4)\right\} \ ,$$

$$\mathrm{a_c} = \max(\mathrm{R}, 4d/\rho, 1) \ , \theta_{\mathrm{c}} = \rho^2/8 \ , \omega_{\mathrm{c}} = \omega(2^{-1}, (8/\rho)\log(4\theta_{\mathrm{c}}^{-1}\beta_{\mathrm{c}})) \ .$$

Proof. Under **H**2(V), [Durmus and Moulines, 2015, Proposition 1] shows that (34) holds for $\mathrm{W_c}$ with constants θ_{c} and β_{c}. Using that for all $a_1, a_2 \in \mathbb{R}$, $\mathrm{e}^{(a_1+a_2)/2} \leq (1/2)(\mathrm{e}^{a_1} + \mathrm{e}^{a_2})$, **G** 2-(ii) holds for $\delta = 2\theta_{\mathrm{c}}^{-1}\beta_{\mathrm{c}}$ and $\mathrm{R} = (8/\rho)\log(4\theta_{\mathrm{c}}^{-1}\beta_{\mathrm{c}})$. As a consequence (51) follows from Theorem 9 with $\epsilon = 1/2$. In addition, [Meyn and Tweedie, 1993, Theorem 4.3-(ii)], (34) implies that $\int_{\mathbb{R}^d} \mathrm{W_c}(y)\mu(\mathrm{d}y) \leq \beta_{\mathrm{c}}\theta_{\mathrm{c}}^{-1}$. The proof of (52) is then concluded using this bound and (51). □

Theorem 13. *Assume **H**2(V) and **H**3(V). Then for all $t \geq 0$, $x, y \in \mathbb{R}^d$ we have*

$$\|\mathbf{P}_t^L(x,\cdot) - \mathbf{P}_t^L(y,\cdot)\|_{\mathrm{TV}} \leq \mathrm{C_s}\left\{1 + \|x - x^\star\| + \|y - x^\star\|\right\}\kappa_{\mathrm{s}}^{2t}$$

$$\|\mathbf{P}_t^L(x,\cdot) - \mu\|_{\mathrm{TV}} \leq \mathrm{C_s}\left\{1 + \|x - x^\star\| + (d/(2m_{\mathrm{s}}) + \mathrm{R_s})^{1/2}\right\}\kappa_{\mathrm{s}}^{2t} \ ,$$

where κ_{s} is defined by (46) and $\mathrm{C_s} = 5/4$.

Proof. The first bound is straightforward application of Theorem 11 and the triangle inequality. For the second one, integrating with respect to μ implies that

$$\|\mathbf{P}_t^L(x,\cdot)-\mu\|_{\mathrm{TV}} \leq \left\{5/4+\|x-x^\star\|+\int_{\mathbb{R}^d}\|y-x^\star\|\,\mathrm{d}\mu(y)+(1+\mathrm{R_s})\mathrm{e}^{m_\mathrm{s}\omega(2^{-1},\mathrm{R_s})/2}\right\}\kappa_\mathrm{s}^{2t}\,.$$

It remains to show that $\int_{\mathbb{R}^d}\|y-x^\star\|\,\mathrm{d}\mu(y)\leq((d/2m_\mathrm{s})+\mathrm{R_s})^{1/2}$. For this, we establish a drift inequality for the generator $\mathscr{A}^L$ of the Langevin SDE associated with . Consider the function $\mathsf{W_s}(x)=\|x-x^\star\|^2$. For all $x\in\mathbb{R}^d$, we have using $\nabla F(x^\star)=0$,

$$\mathscr{A}^L\mathsf{W_s}(x)=-\langle\nabla U(x)-\nabla U(x^\star),x-x^\star\rangle+d\,.$$

Therefore by **G**3, for all $x\in\mathbb{R}^d$, $\|x-x^\star\|\geq\mathrm{R_s}$, we get

$$\mathscr{A}^L\mathsf{W_s}(x)=-2m_\mathrm{s}\mathsf{W_s}(x)+d\,,$$

and for all $x\in\mathbb{R}^d$,

$$\mathscr{A}^L\mathsf{W_s}(x)=-2m_\mathrm{s}\mathsf{W_s}(x)+d+2m_\mathrm{s}\mathrm{R_s}\,.$$

By [Meyn and Tweedie, 1993, Theorem 4.3-(ii)], we get $\int_{\mathbb{R}^d}\mathsf{W_s}(y)\mathrm{d}\mu(y)\leq d+m_\mathrm{s}\mathrm{R_s}$, and the Cauchy-Schwarz inequality concludes the proof. □

8 Drift inequalities for ULA

We now study the stability of $\{X_k^\mathrm{M},\ k\in\mathbb{N}\}$ defined by (3) under **H**2(V). For this purpose, we establish a geometric drift condition for R_γ with $\gamma>0$.

Proposition 14. *Assume **H**2(V). Then for all $\gamma\in\left(0,L_\mathrm{c}^{-1}\right]$ and $x\in\mathbb{R}^d$,*

$$R_\gamma\mathrm{W_c}(x)\leq\lambda_\mathrm{c}^\gamma\mathrm{W_c}(x)+\left(\mathrm{e}^{(\rho_\mathrm{c}\gamma/4)(d+(\rho_\mathrm{c}/8))}-\lambda_\mathrm{c}^\gamma\right)\mathrm{e}^{\rho_\mathrm{c}(M_\mathrm{c}^2+1)^{1/2}/4}\mathbb{1}_{\mathrm{B}(x^\star,M_\mathrm{c})}(x)\,,$$

where $M_\mathrm{c}=\max(1,2d/\rho_\mathrm{c},\mathrm{R_c})$ and

$$\lambda_\mathrm{c}=\mathrm{e}^{-2^{-4}\rho_\mathrm{c}^2(2^{1/2}-1)}\,. \tag{53}$$

Proof. Set $\alpha=\rho_\mathrm{c}/4$ and for all $x\in\mathbb{R}^d$, $\mathrm{f}(x)=(\|x-x^\star\|^2+1)^{1/2}$. Since f is 1-Lipschitz, we have by the log-Sobolev inequality [Boucheron et al., 2013, Theorem 5.5] for all $x\in\mathbb{R}^d$,

$$R_\gamma\mathrm{W_c}(x)\leq\mathrm{e}^{\alpha R_\gamma\mathrm{f}(x)+\alpha^2\gamma}\leq\mathrm{e}^{\alpha\sqrt{\|x-\gamma\nabla V(x)-x^\star\|^2+2\gamma d+1}+\alpha^2\gamma}\,. \tag{54}$$

Under **H**2 since $x^\star$ is a minimizer of V, [Nesterov, 2004, Theorem 2.1.5 Equation (2.1.7)] shows that for all $x\in\mathbb{R}^d$,

$$\langle\nabla V(x),x-x^\star\rangle\geq(2L_\mathrm{c})^{-1}\|\nabla V(x)\|^2+\rho_\mathrm{c}\|x-x^\star\|\mathbb{1}_{\{\|x-x^\star\|\geq\mathrm{R_c}\}}\,,$$

which implies that for all $x\in\mathbb{R}^d$ and $\gamma\in\left(0,L_\mathrm{c}^{-1}\right)$, we have

$$\|x-\gamma\nabla V(x)-x^\star\|^2\leq\|x-x^\star\|^2-2\gamma\rho_\mathrm{c}\|x-x^\star\|\mathbb{1}_{\{\|x-x^\star\|\geq\mathrm{R_c}\}}\,. \tag{55}$$

Using this inequality and for all $u \in [0,1]$, $(1-u)^{1/2} - 1 \leq -u/2$, we have for all $x \in \mathbb{R}^d$, satisfying $\|x - x^\star\| \geq M_{\mathrm{c}} = \max(1, 2d\rho_{\mathrm{c}}^{-1}, \mathrm{R}_{\mathrm{c}})$,

$$\begin{aligned}\left(\|x - \gamma\nabla V(x) - x^\star\|^2 + 2\gamma d + 1\right)^{1/2} - \mathrm{f}(x) &\leq \mathrm{f}(x)\left\{\left(1 - 2\gamma \mathrm{f}^{-2}(x)(\rho_{\mathrm{c}}\|x - x^\star\| - d)\right)^{1/2} - 1\right\} \\ &\leq -\gamma \mathrm{f}^{-1}(x)(\rho_{\mathrm{c}}\|x - x^\star\| - d) \leq -(\rho_{\mathrm{c}}\gamma/2)\|x - x^\star\|\,\mathrm{f}^{-1}(x) \leq -2^{-3/2}\rho_{\mathrm{c}}\gamma\,.\end{aligned}$$

Combining this inequality and (54), we get for all $x \in \mathbb{R}^d$, $\|x - x^\star\| \geq M_{\mathrm{c}}$,

$$R_\gamma \mathrm{W}_{\mathrm{c}}(x)/\mathrm{W}_{\mathrm{c}}(x) \leq \mathrm{e}^{\gamma\alpha(\alpha - 2^{-3/2}\rho_{\mathrm{c}})} = \lambda_{\mathrm{c}}\,.$$

By (55) and and the inequality for all $a, b \geq 0$, $\sqrt{a+1+b} - \sqrt{1+b} \leq a/2$, we get for all $x \in \mathbb{R}^d$,

$$\sqrt{\|x - \gamma\nabla V(x) - x^\star\|^2 + 2\gamma d + 1} - \mathrm{f}(x) \leq \gamma d\,.$$

Then using this inequality in (54) concludes the proof. □

Corollary 15. *Assume **H**2(V). Let $\{\gamma_k,\ k \in \mathbb{N}^*\}$ be a nonincreasing sequence with $\gamma_1 \leq L_{\mathrm{c}}^{-1}$.*

a) For all $n \geq 0$ and $x \in \mathbb{R}^d$

$$Q_\gamma^n \mathrm{W}_{\mathrm{c}}(x) \leq \lambda_{\mathrm{c}}^{\Gamma_{1,n}} \mathrm{W}_{\mathrm{c}}(x) + \mathrm{c}_{\mathrm{c}}(1 - \lambda_{\mathrm{c}}^{\Gamma_{1,n}})/(1 - \lambda_{\mathrm{c}}^{\gamma_1})\,, \tag{56}$$

where

$$\mathrm{c}_{\mathrm{c}} = \gamma_1\left\{(\rho_{\mathrm{c}}/4)(d + (\rho_{\mathrm{c}}\gamma_1/4)) - \log(\lambda_{\mathrm{c}})\right\}\mathrm{e}^{\rho_{\mathrm{c}}(M_{\mathrm{c}}+1)/4 + (\rho_{\mathrm{c}}\gamma_1/4)(d + (\rho_{\mathrm{c}}\gamma_1/4))}\,. \tag{57}$$

b) For all $n \geq 0$ and $x \in \mathbb{R}^d$,

$$\int_{\mathbb{R}^d} \|y - x^\star\|^2 Q_\gamma^n(x, \mathrm{d}y) \leq \left\{4\rho_{\mathrm{c}}^{-1}\left(1 + \log\left\{\lambda_{\mathrm{c}}^{\Gamma_{1,n}}\mathrm{W}_{\mathrm{c}}(x) + \mathrm{c}_{\mathrm{c}}\frac{1 - \lambda_{\mathrm{c}}^{\Gamma_{1,n}}}{1 - \lambda_{\mathrm{c}}^{\gamma_1}}\right\}\right)\right\}^2\,,$$

Proof. a) By Proposition 14 and the inequality for all $t \geq 0$, $\mathrm{e}^t - 1 \leq t\mathrm{e}^t$, we have that for all $x \in \mathbb{R}^d$ and $\gamma \leq \gamma_1$,

$$R_\gamma \mathrm{W}_{\mathrm{c}}(x) \leq \lambda_{\mathrm{c}}^\gamma \mathrm{W}_{\mathrm{c}}(x) + \mathrm{c}_{\mathrm{c}}\gamma\,.$$

By a straightforward induction, we get for all $n \geq 0$ and $x \in \mathbb{R}^d$,

$$Q_\gamma^n \mathrm{W}_{\mathrm{c}}(x) \leq \lambda_{\mathrm{c}}^{\Gamma_{1,n}} \mathrm{W}_{\mathrm{c}}(x) + \mathrm{c}_{\mathrm{c}} \sum_{i=1}^n \gamma_i \lambda_{\mathrm{c}}^{\Gamma_{i+1,n}}\,. \tag{58}$$

Note that for all $n \geq 1$, we have

$$(1 - \lambda_{\mathrm{c}}^{\gamma_1}) \sum_{i=1}^n \gamma_i \lambda_{\mathrm{c}}^{\Gamma_{i+1,n}} = \sum_{i=1}^n \gamma_i \lambda_{\mathrm{c}}^{\Gamma_{i+1,n}} - \sum_{i=1}^n \gamma_i \lambda_{\mathrm{c}}^{\Gamma_{i,n}} \leq \gamma_1 \sum_{i=1}^n (\lambda_{\mathrm{c}}^{\Gamma_{i+1,n}} - \lambda_{\mathrm{c}}^{\Gamma_{i,n}}) \leq \gamma_1(1 - \lambda_{\mathrm{c}}^{\Gamma_{1,n}})\,.$$

The proof is then completed using this inequality in (58).

b) Let $n \geq 0$ and $x \in \mathbb{R}^d$. Consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by for all $t \in \mathbb{R}$, $h(t) = \exp\left\{(\rho/4)(t + (4/\rho)^2)^{1/2}\right\}$. Since this function is convex on $\mathbb{R}_+$, we have by the Jensen inequality and the inequality for all $t \geq 0$, $h(t) \leq \mathrm{e}^{1+(\rho/4)(t+1)^{1/2}}$,

$$h\left(\int_{\mathbb{R}^d} \|y - x^\star\|^2 \, Q_\gamma^n(x, \mathrm{d}y)\right) \leq \mathrm{e}^1 Q_\gamma^n \mathrm{W}_\mathrm{c}(x) \; .$$

The proof is then completed using (56) and that h is ono-to-one with for all $t \geq 1$, $h^{-1}(t) \leq \left(4\rho^{-1}\log(t)\right)^2$.

□

Assuming the additional assumption **H** 3(V), an other drift inequality is derived in the following Proposition.

Proposition 16. *Assume **H**2(V) and **H**3(V). Then for all $\gamma \in \left(0, 4m_\mathrm{s}L_\mathrm{c}^{-1}\right)$ and $x \in \mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} \|y - x^\star\|^2 \, R_\gamma(x, \mathrm{d}y) \leq \lambda_\mathrm{s} \|x - x^\star\|^2 + (2\gamma d + (\mathrm{R}_\mathrm{s}\gamma L_\mathrm{c})^2) \; ,$$

where

$$\lambda_\mathrm{s} = (1 - \gamma(2m_\mathrm{s} - \gamma L_\mathrm{c})) \; . \tag{59}$$

Proof. For all $x \in \mathbb{R}^d$, we have by $\nabla V(x^\star) = 0$ and **H**2(V) and **H**3(V)

$$\begin{aligned}\int_{\mathbb{R}^d} \|y - x^\star\|^2 \, R_\gamma(x, \mathrm{d}y) &= \|x - x^\star + \gamma(\nabla V(x^\star) - \nabla V(x))\|^2 + 2\gamma d \\ &\leq (1 + (L_\mathrm{c}\gamma)^2) \|x - x^\star\|^2 - 2\gamma \langle \nabla V(x) - \nabla V(x^\star), x - x^\star \rangle + 2\gamma d \; . \end{aligned} \tag{60}$$

Then for all $x \in \mathbb{R}^d$, $\|x - x^\star\| \geq \mathrm{R}_\mathrm{s}$, we get

$$\int_{\mathbb{R}^d} \|y - x^\star\|^2 \, R_\gamma(x, \mathrm{d}y) \leq \lambda_\mathrm{s} \|x - x^\star\|^2 + 2\gamma d \; .$$

Using again (60) and **H**2(V), it yields for all $x \in \mathbb{R}^d$, $\|x - x^\star\| \leq \mathrm{R}_\mathrm{s}$,

$$\int_{\mathbb{R}^d} \|y - x^\star\|^2 \, R_\gamma(x, \mathrm{d}y) \leq \mathrm{c}_\mathrm{s} \; ,$$

which concludes the proof. □

Corollary 17. *Assume **H**2(V). Let $\{\gamma_k, \; k \in \mathbb{N}^*\}$ be a nonincreasing sequence with $\gamma_1 < 4m_\mathrm{s}L_\mathrm{c}^{-1}$. For all $n \geq 0$ and $x \in \mathbb{R}^d$,*

$$\int_{\mathbb{R}^d} \|y - x^\star\|^2 \, Q_\gamma^n(x, \mathrm{d}y) \leq \lambda_\mathrm{s}^{\Gamma_{1,n}} \|x - x^\star\|^2 + \mathrm{c}_\mathrm{s}(1 - \lambda_\mathrm{s}^{\Gamma_{1,n}})/(1 - \lambda_\mathrm{s}^{\gamma_1}) \; ,$$

where

$$\mathrm{c}_\mathrm{s} = \gamma_1(2d + \gamma_1(\mathrm{R}_\mathrm{s}L_\mathrm{c})^2) \; . \tag{61}$$

Proof. The proof is the same as the one of Corollary 15 and is omitted. □

8.1 Proof of Theorem 3

We bound the two terms of the decomposition given in (17). By Theorem 12 and Corollary 15-a), we get

$$\|\mu_0 Q_\gamma^n \mathbf{P}^L_{\Gamma_{n+1,p}} - \mu\|_{\mathrm{TV}} \leq \mathrm{C_c} \kappa_{\mathrm{c}}^{\Gamma_{n+1,p}} \left\{ \beta_{\mathrm{c}} \theta_{\mathrm{c}}^{-1} + \lambda_{\mathrm{c}}^{\Gamma_{1,n}} \mathrm{W_c}(x) + \mathrm{c_c}(1 - \lambda_{\mathrm{c}}^{\Gamma_{1,n}})/(1 - \lambda_{\mathrm{c}}^{\gamma_1}) \right\} .$$

It remains to bound the second term. Using [Durmus and Moulines, 2015, Section 3, Eq. 24], it holds

$$\|\delta_x Q_\gamma^{n+1,p} - \delta_x P_{\Gamma_{n+1,p}}\|_{\mathrm{TV}}^2 \leq 2^{-1} L_{\mathrm{c}}^2 \sum_{k=n}^{p-1} \left\{ (\gamma_{k+1}^3/3) \int_{\mathbb{R}^d} \|\nabla V(y)\|^2 \, Q_\gamma^k(x, \mathrm{d}y) + d\gamma_{k+1}^2 \right\} . \quad (62)$$

Using $\nabla V(x^\star) = 0$, $\mathbf{H}2(V)$ and Corollary 15-b), we have for all $k \in \{n, \ldots, p\}$,

$$\int_{\mathbb{R}^d} \|\nabla V(y)\|^2 \, Q_\gamma^k(x, \mathrm{d}y) \leq L_{\mathrm{c}}^2 \left(4\rho_{\mathrm{c}}^{-1} \left\{ 1 + \log \left\{ \mathrm{W_c}(x) + \mathrm{c_c} \, (1 - \lambda_{\mathrm{c}}^{\gamma_1})^{-1} \right\} \right\} \right)^2 .$$

Using this inequality in (62) concludes the proof.

8.2 Proof of Theorem 4

The proof is the same as the one of Theorem 3 using Theorem 13 instead of Theorem 12 and Corollary 17 instead of Corollary 15.

8.3 Proof of Proposition 5

Under $\mathbf{H}2$, R_γ is irreducible with respect to the Lebesgue measure and weak Feller, which implies by [Meyn and Tweedie, 2009, Proposition 6.2.8] that every compact set is small. Using [Meyn and Tweedie, 2009, Theorem 14.0.1] and Proposition 14, R_γ admits a unique invariant probability measure π_γ. By Theorem 3, there exists some constant C_1 and C_2 such that

$$\|R_\gamma^p - \pi_\gamma\|_{\mathrm{TV}} \leq (C_1 \kappa_{\mathrm{c}}^{\gamma p} + C_2 \gamma p^{1/2}) \mathrm{W_c}(x) . \quad (63)$$

Using again [Meyn and Tweedie, 2009, Theorem 14.0.1] and Proposition 14, $\pi_\gamma(\mathrm{W_c}) < +\infty$. So integrating (63) with respect to π_γ and choosing $p = \max(1, -\log(\gamma)/\gamma)$ concludes the proof.

References

Afonso, M. V., Bioucas-Dias, J. M., and Figueiredo, M. A. (2011). An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems. *Trans. Img. Proc.*, 20(3):681–695.

Bakry, D., Barthe, F., Cattiaux, P., and Guillin, A. (2008). A simple proof of the Poincaré inequality for a large class of probability measures. *Electronic Communications in Probability [electronic only]*, 13:60–66.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities*. Oxford University Press, Oxford. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

Chambolle, A. (2004). An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1):89–97.

Chen, M. F. and Li, S. F. (1989). Coupling methods for multidimensional diffusion processes. *Ann. Probab.*, 17(1):151–177.

Dalalyan, A. (2014). Theoretical guarantees for approximate sampling from a smooth and log-concave density. submitted 1412.7392, arXiv.

Durmus, A. and Moulines, E. (2015). Non-asymptotic convergence analysis for the unadjusted langevin algorithm. submitted 1507.05021, arXiv.

Eberle, A. (2015). Reflection couplings and contraction rates for diffusions. *Probab. Theory Related Fields*, pages 1–36.

Ermak, D. L. (1975). A computer simulation of charged particles in solution. i. technique and equilibrium properties. *The Journal of Chemical Physics*, 62(10):4189–4196.

Ethier, S. N. and Kurtz, T. G. (1986). *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York. Characterization and convergence.

Florenzano, M. and Le Van, C. (2001). *Finite dimensional convexity and optimization*, volume 13 of *Studies in Economic Theory*. Springer-Verlag, Berlin. In cooperation with Pascal Gourdel.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862.

Grenander, U. (1983). Tutorial in pattern theory. Division of Applied Mathematics, Brown University, Providence.

Grenander, U. and Miller, M. I. (1994). Representations of knowledge in complex systems. *J. Roy. Statist. Soc. Ser. B*, 56(4):549–603. With discussion and a reply by the authors.

Ikeda, N. and Watanabe, S. (1989). *Stochastic Differential Equations and Diffusion Processes*. North-Holland Mathematical Library. Elsevier Science.

Kampf, J. (2009). On weighted parallel volumes. *Beiträge Algebra Geom*, 50(2):495–519.

Khas'minskii, R. Z. (1960). Ergodic properties of recurrent diffusion processes and stabilization of the solution to the cauchy problem for parabolic equations. *Theory of Probability & Its Applications*, 5(2):179–196.

Lamberton, D. and Pagès, G. (2002). Recursive computation of the invariant distribution of a diffusion. *Bernoulli*, 8(3):367–405.

Lemaire, V. (2005). *Estimation de la mesure invariante d'un processus de diffusion*. PhD thesis, Université Paris-Est.

Lindvall, T. and Rogers, L. C. G. (1986). Coupling of multidimensional diffusions by reflection. *Ann. Probab.*, 14(3):860–872.

Meyn, S. and Tweedie, R. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition.

Meyn, S. P. and Tweedie, R. L. (1993). Stability of Markovian processes. III. Foster-Lyapunov criteria for continuous-time processes. *Adv. in Appl. Probab.*, 25(3):518–548.

Neal, R. M. (1993). Bayesian learning via stochastic dynamics. In *Advances in Neural Information Processing Systems 5, [NIPS Conference]*, pages 475–482, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Applied Optimization. Springer.

Oliveira, J. P., Bioucas-Dias, J. M., and Figueiredo, M. A. (2009). Adaptive total variation image deblurring: A majorization–minimization approach. *Signal Processing*, 89(9):1683 – 1693.

Parikh, N. and Boyd, S. (2013). *Proximal Algorithms*. Foundations and Trends(r) in Optimization. Now Publishers.

Parisi, G. (1981). Correlation functions and computer simulations. *Nuclear Physics B*, 180:378–384.

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686.

Pereyra, M. (2015). Proximal markov chain monte carlo algorithms. *Statistics and Computing*, pages 1–16.

Polson, N. G., Scott, J. G., and Willard, B. T. (2015). Proximal algorithms in statistics and machine learning. *Statist. Sci.*, 30(4):559–581.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Rossky, P. J., Doll, J. D., and Friedman, H. L. (1978). Brownian dynamics as smart Monte Carlo simulation. *The Journal of Chemical Physics*, 69(10):4628–4633.

Schneider, R. (2013). *Convex bodies: the Brunn–Minkowski theory*. Number 151. Cambridge University Press.